Data Analysis

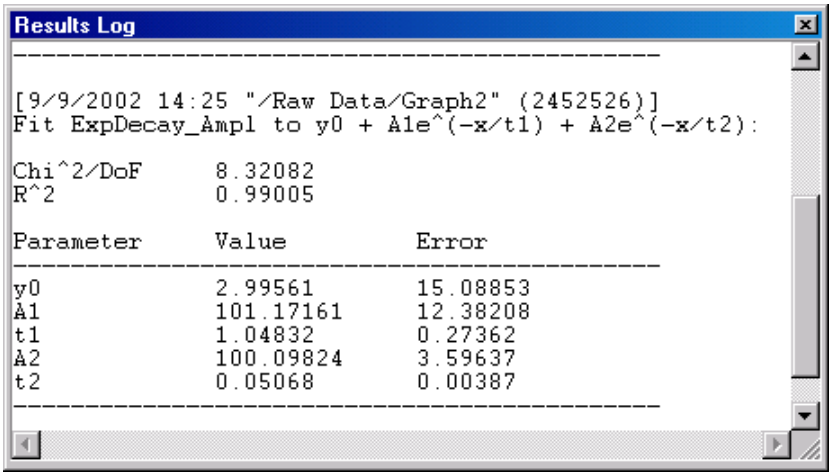
Selecting the Active Data Plot for Analysis

When performing analysis on a data plot in a graph window, you must make the data plot you want to analyze the *active* data plot. To make a data plot active, select the data plot from the data list at the bottom of the **Data** menu. The data list includes all data plots in the active layer. The currently active data plot is checked.

Many analysis menu commands and tool functions, when applied to a data plot, will also make changes to the associated worksheet data set(s). In addition, some analysis menu commands and tool functions add new data sets to existing worksheets. The **Subtract** and **Smoothing** menu commands fall into this category.

The Results Log

Origin automatically routes results from most commands on the **Analysis** menu as well as results from the Baseline, Linear Fit, Polynomial Fit, and Sigmoidal Fit tools to the Results Log. Each entry in the Results Log includes a date/time stamp, the project file location, the data set, the type of analysis performed, and the results.



The Results Log can display as a floating window, or it can be docked to the edge of the workspace. If the Results Log is closed, it can be re-opened by selecting **View:Results Log** or by clicking the Results Log button on the Standard toolbar.

Note: To prevent the Results Log from docking when you move it in the workspace, press CTRL while dragging the window to the new location.

Origin allows you to view all your project results in the Results Log, or only those analysis results from your current Project Explorer folder, or subfolders. To control the view, select **View:View Mode:*View Mode***, or right-click in the Results Log and select the view mode from the shortcut menu.

You cannot edit the contents of the Results Log. However, you can clear the Results Log by right-clicking in the Results Log and selecting **Clear** from the shortcut menu. Additionally, you can delete the last entry in the Results Log by selecting **Clear Last Entry** from the shortcut menu. If the Results Log view mode is set to either **View Results in Active Folder** or **View Results in Active Folder & Subfolders** (both available from the shortcut menu), then **Clear Last Entry** will delete the last visible entry for the current view mode.

To undo the last cleared entry, select **Restore Last Cleared Entry** from the shortcut menu.

You can also print the visible entries in the Results Log by selecting **Print** from the shortcut menu.

Data Exploration Tools

Origin offers a number of tools to select a region of a data plot, view coordinate values, and enhance the data plot display. These tools are available from the Tools toolbar (**View:Toolbars**).

The Data Display Tool

Origin's Data Display tool simulates a display panel, providing a dynamic display of the selected data point's, or the screen's, XY coordinates. The Data Display tool opens when the Screen Reader, Data Reader, Data Selector, or Draw tool from the Tools toolbar are selected. Additionally, the tool opens when moving or removing data points, or when translating your data.

Resize the tool to enlarge or reduce its display. As a floating tool, position it anywhere in Origin's workspace. To dock the Data Display tool, right-click on the tool and select (check) **Docking View** from the shortcut menu. The tool automatically docks to the toolbar region. If the **Docking View** shortcut menu command was already checked, double-click on the tool to dock it to the toolbar region. Alternatively, drag the tool to the edge of the workspace. The docked tool remains available throughout your Origin session, as well as for future Origin sessions.

The Shortcut Menus Available from the Data Display Tool

Shortcut Menu Commands Available from the Tool's Title Bar

Move: Activates the tool so that it can be moved using the arrow keyboard keys.

Hide: Hides the tool for the current data exploration routine.

Shortcut Menu Commands Available from the Tool

Fixed Size: This shortcut menu command is only available when the Automatically Fit to Display check box in the Data Display Format dialog box is cleared.

Fixes the size of the text in the tool to the current size when this menu command is selected. Resizing the tool horizontally or vertically has no affect on the text size.

Fit Horizontally: This shortcut menu command is only available when the Automatically Fit to Display check box in the Data Display Format dialog box is cleared.

Displays the text in the tool so that it always displays in the full width of the tool. Resizing the tool horizontally causes the text size to change accordingly. Resizing the tool vertically has no affect on the text display.

Fit Vertically: This shortcut menu command is only available when the Automatically Fit to Display check box in the Data Display Format dialog box is cleared.

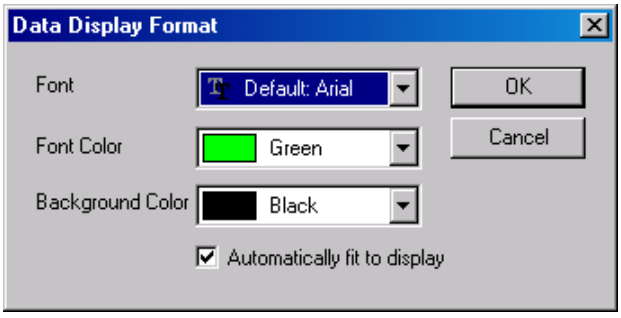
Displays the text in the tool so that it always displays in the full height of the tool. Resizing the tool vertically causes the text size to change accordingly. Resizing the tool horizontally has no affect on the text display.

Docking View: Automatically docks the tool to the toolbar region (if the shortcut menu command was not already selected). If the docked tool is then dragged to a floating tool, double-click on the tool to restore its docked position. Alternatively, drag the tool to the edge of the workspace.

Properties: Opens the Data Display Format dialog box.

Customizing the Display of the Data Display Tool

To customize the tool's coordinate display including the font, font color, and tool background color, right-click on the tool and select **Properties** from the shortcut menu. This shortcut menu command opens the Data Display Format dialog box.



The Data Display Format Dialog Box

The Font Drop-down List

Select the tool's text font display from this drop-down list.

The Font Color Drop-down List

Select the tool's text color display from this drop-down list.

The Background Color Drop-down List

Select the tool's background color display from this drop-down list.

The Automatically Fit to Display Check Box

Select this check box to ensure that all the text within the tool displays when the tool is resized. The text font size changes to accommodate changes in tool size.

When this check box is cleared, the **Fixed Size**, **Fit Horizontally**, and **Fit Vertically** shortcut menu commands are available. Select the desired command to control the text display in the tool.

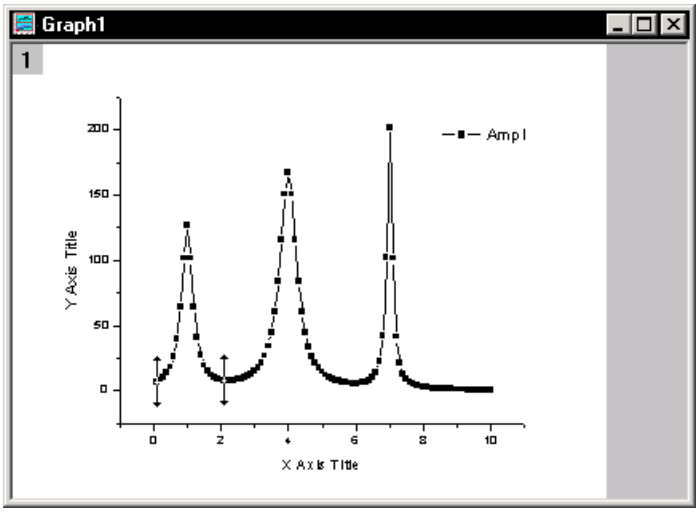
The Data Selector Tool

To select a range of a data plot for analysis, click the Data Selector tool on the Tools toolbar. Data markers display at both ends of the active data plot. Additionally, the Data Display tool opens if it is not already open.

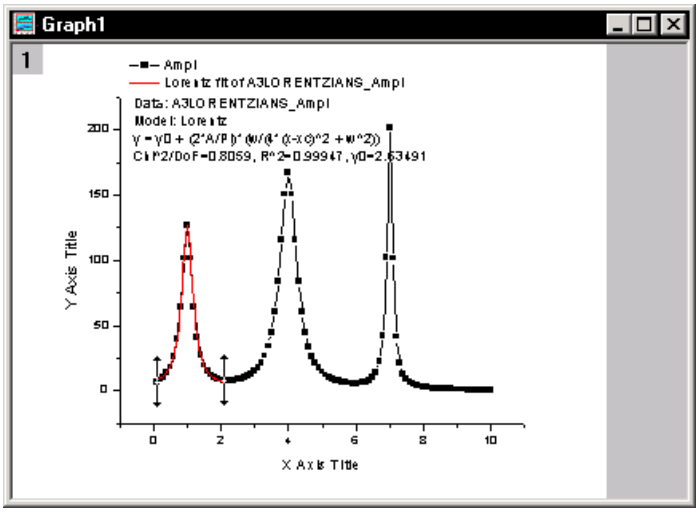
To mark the data segment of interest, click and drag the markers with the mouse. You can also use the left and right arrow keys to select a marker. The CTRL + left or right arrow keys move the selected marker to the next data point. Holding both the SHIFT and CTRL keys while depressing the left or right arrow keys moves the data markers in increments of five along the data plot. (Note: If your X data is not sorted, you may need to sort the data before selecting a range. To do this, activate the worksheet and select **Analysis:Sort Worksheet:Ascending**.)

As with the Data Reader tool, you can press the spacebar to increase the cross hair size.

After you have defined the range of interest, press ESC or click the Pointer button on the Tools toolbar.



Any analysis operations you perform on this data plot will apply to the selected region only.

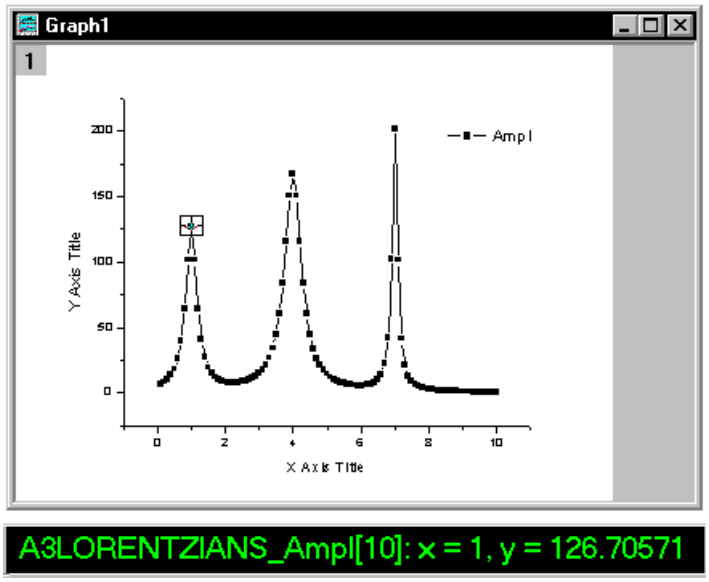


To hide the data outside this range, select **Data:Set Display Range**.

To remove the selection range, select **Data:Reset to Full Range**.

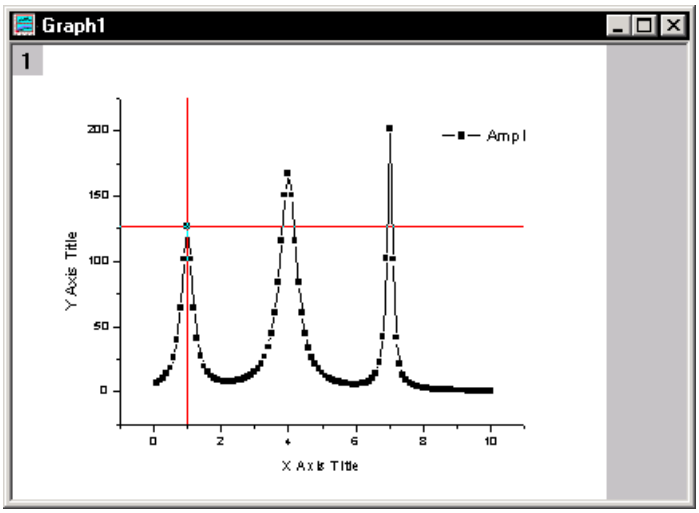
The Data Reader Tool

To read the X, Y, and Z (for 3D and contour) values for a data point, click the Data Reader button on the Tools toolbar. This action opens the Data Display tool if it is not already open. Click on the desired data point to read its X,Y, and Z coordinates in the Data Display tool.



To move the cross-hair to the next data point along the data plot, use the left and right arrow keys or click on the data point.

To change the vertical and horizontal cross hair size after clicking on a point, press the spacebar. Continue pressing the spacebar to further increase the size of the cross hairs.



Press ESC or click the Pointer button on the Tools toolbar when you are finished.

A Note on Improving the Search

When you click on a data plot using the Data Reader tool, Origin begins searching for the nearest point by one of two methods based on the number of points in the data plot and the Bisection Search Points combination box value on the Miscellaneous tab of the Options dialog box (**Tools:Options**).

1) When the number of points is less than the bisection search value, then sequential search is used.

2) When the number of points is greater than the bisection search value, then bisection search is used.

Bisection search is designed to find the nearest point rapidly, but it only works if the X data set is sorted. The middle point of the data set is tested first to determine which half contains the nearest point. This half becomes the new search set. If the X data set is not sorted, then this search scheme fails and you get a 'beep error'. (Actually a few points will be found this way purely by chance.)

If a search fails because bisection searching is being used on an unsorted (in X) data set, then you can either:

1) Sort the source worksheet on the X column.

2) Set the Bisection Search Points combination box to a value larger than the number of data points to 'turn off' bisection search.

The Screen Reader Tool

To read the X, Y, and Z (for 3D and contour) values for any point on the screen, click the Screen Reader button on the Tools toolbar. This action opens the Data Display tool if it is not already open. Click on the desired screen location to read its X,Y, and Z coordinates in the Data Display tool.

As with the Data Reader tool, you can press the spacebar to increase the cross hair size.

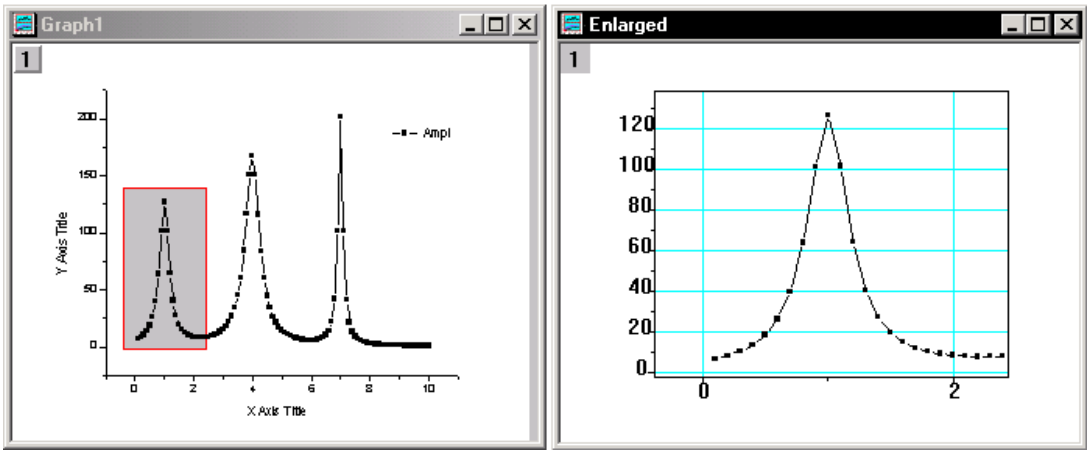
Press ESC or click the Pointer button on the Tools toolbar when you are finished.

The Enlarger Tool and Undo Enlarge

To magnify a portion of a data plot, click the Enlarger tool on the Tools toolbar. The close-up of the data can be viewed in the current window or in a new window. When a data plot is magnified using the Enlarger tool, the axes are rescaled to show an expanded view of the data.

=> To view the magnified data in the current window, drag the area of interest in the graph. To re-display the entire data plot, click the Undo Enlarge button .

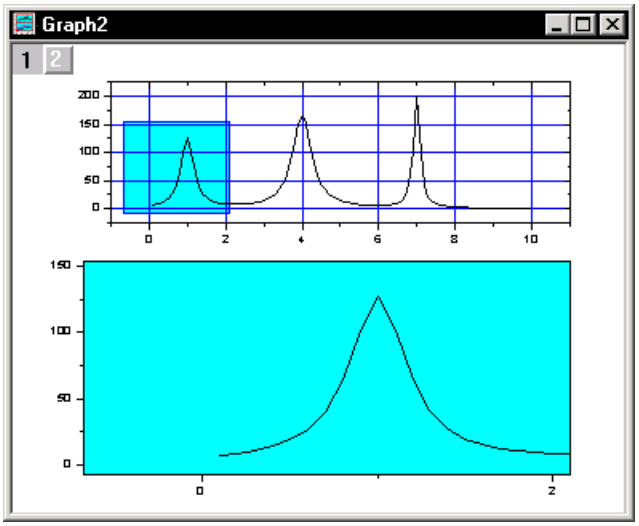
=> To view the magnified data in a new window, hold down the CTRL key while dragging the area of interest in the graph window. Release the mouse button, then release the CTRL key. Origin opens a new graph window named Enlarged. A rectangle object with sizing handles appears in the original graph window (if the sizing handles are not displayed, click on the object to select it).



To change the data segment in the Enlarged window, resize the rectangular object or change its position by dragging in the original graph window. The Enlarged window updates accordingly.

After you finish analyzing the data, you can click on the rectangle object and delete it.

=> To view the magnified data in the same graph window (along with the full range of data), highlight the desired worksheet column(s) and click the Zoom button on the 2D Graphs Extended toolbar. A graph window with two layers opens. The top layer displays the entire data range and the bottom layer provides a zoomed-in view of your data plot.

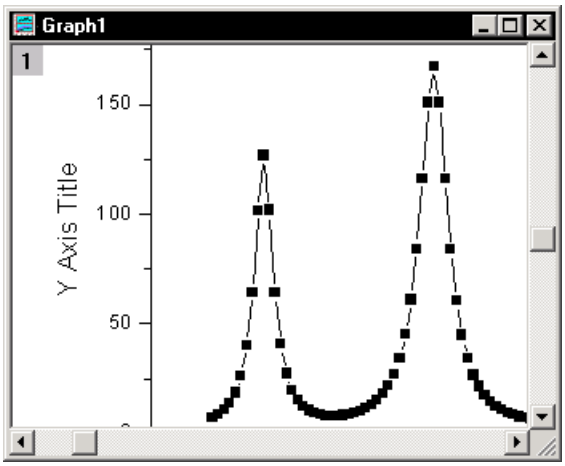


To change the data segment in the bottom layer, resize the rectangular object or change its position by dragging in the top layer.

After you finish analyzing the data, you can click on the rectangle object and delete it.

Zoom In and Zoom Out

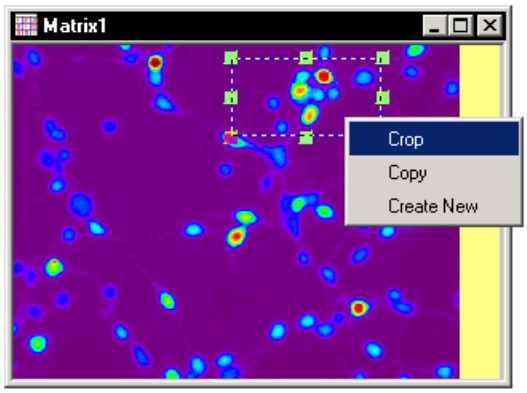
To display a close-up view of the graph page, click the Zoom In button on the Graph toolbar, then click on the desired zoom location in the graph window. Origin zooms the page, centering the zoom view at this location in the window (the axes are not rescaled).



To zoom in closer, click the button again. To zoom out, click the Zoom Out button. To return the view to the full page, click the Whole Page button.

Region of Interest (Image Data)

When you import image data in a matrix, you can select a region of the image using the Rectangle Tool on the Tools toolbar. (When a matrix with an image is active, the Rectangle tool displays in the "region of interest mode" by default. This setting is controlled from the **Tools:Region of Interest Tools** menu command.) Select the Rectangle Tool and then drag your region of interest. Then right-click in this region and select **Crop**, **Copy**, or **Create New** from the shortcut menu.



Masking Data from Analysis

The Mask toolbar (**View:Toolbars**) allows you to exclude ranges of data or individual data points from analysis. It is most useful if you want to analyze only a specific section of your data, or if you have erroneous data points that you do not want included in your analysis. The Mask toolbar is available for both worksheet and graph windows.

Note: The following **Analysis** menu commands are not affected by masking data (masking is ignored): Simple Math, FFT Filter, Translate Horizontal/Vertical, Subtract Reference Data/Straight Line (graph is active), and Normalize (worksheet is active).

To Mask Data in a Worksheet

When a worksheet is active, select a range of data or a single cell and click Mask Range . The cell color will change to red and become masked. The masked data will not be included in any fitting or analysis you do to your data. Once the data is masked, you can change the color of the masked cells by clicking Mask Color . Each time you click this button the color increments through the color palette.

Disable/Enable Masking toggles the masking on and off.

To Mask Data in a Graph

When a graph is active, you can mask a range of data or a single data point.

To mask a single data point, click Mask Point Toggle . This action activates the Data Reader tool. Double-click on the data point that you want to mask. The Data Reader tool closes and the point changes color to red and is masked. To unmask a data point, click Mask Point Toggle and then double-click on the desired data point.

To mask a range of data in a graph, click Mask Range . Data markers display on both ends of the active data plot. Select the desired data range and then press ENTER. Once the data is masked, you can change the color of the masked cells by clicking Mask Color. Each time you click this button the color increments through the color palette.

To unmask a range of data points, click Unmask Range . You can now select a range to unmask (not necessarily the same range you masked).

Swap Mask removes the masking from the masked data and applies it to the unmasked data.

Hide/Show Masked Points toggles the display of the masked points in the graph.

Disable/Enable Masking toggles the masking on and off.

To learn more about masking data, review the DATA MASKING.OPJ project located in your Origin \SAMPLES\ANALYSIS\DATA MASKING folder.

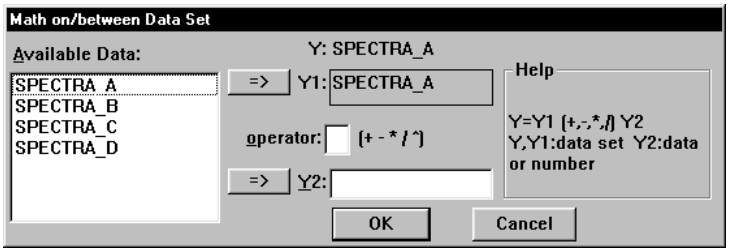
Simple Math

Performing Math Operations

To perform a math operation on or between two data sets, select **Analysis:Simple Math**. This menu command opens the Math on/between Dataset dialog box. Note that the **Analysis:Simple Math** menu command is only available when a graph window is active. When the worksheet is active, similar results can be obtained by selecting **Column:Set Column Values**.

Note: The Simple Math operations will modify worksheet data sets. Additionally, the X data set(s) should be sorted before performing Simple Math operations.

The Math on/between Dataset Dialog Box



1) To perform a math operation, click to select the desired data set from the Available Data list box.

2) Click the upper => button to display this data set in the Y1 text box.

3) Type a numerical value, a variable, or an equation in the Y2 text box. To display a data set in this text box, select the data set from the Available Data list box and click the lower => button.

4) Type an arithmetic operator in the Operator text box.

5) Click OK to perform the math operation.

For example, if you add the **Data1_B** data set into the Y1 text box, the **Data1_C** data set into the Y2 text box, and a subtraction operator (-) in the Operator text box, Origin subtracts each value in **Data1_C** from the corresponding value in **Data1_B** and saves and plots the result as **Data1_B** after clicking OK.

You can also type a numerical constant for the Y2 value. For example, if Y2=35 and the operator is a minus sign (-), then Origin subtracts 35 from each value in the Y1 data set and plots the result after clicking OK.

Note: If the X coordinate points of Y2 do not correspond to the X coordinate points of Y1, Y2 is interpolated or extrapolated for each point in Y1. The interpolation or extrapolation depends on the line connection type for the Y2 data plot. For example, if the line connection type is spline, the interpolation is a spline interpolation.

Performing Calculations on Data Sets Using LabTalk

You can use LabTalk to perform calculations on data sets, either performing the calculations row by row, or using linear interpolation if the data sets have different numbers of elements. In this case, whether or not linear interpolation is used is dependent on your LabTalk notation.

Row by Row Calculations

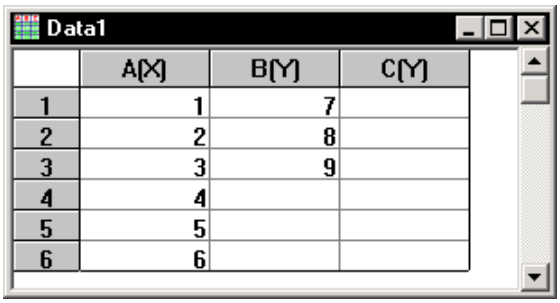
Vector calculations are always performed row by row when you use the general notation:

```
data set = scaler operator data set
```

or

```
data set = data set operator data set
```

This is the case even if the data sets have different numbers of elements. For example, if you start with the following worksheet:



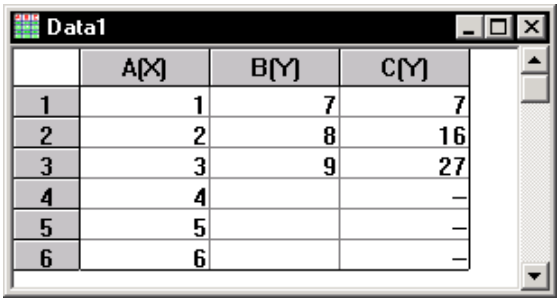
	A(X)	B(Y)	C(Y)
1	1	7	
2	2	8	
3	3	9	
4	4		
5	5		
6	6		

And then type the following LabTalk script in the Script window (**Window:Script Window**):

```
data1_C = data1_A * data1_B
```
(press ENTER)

The elements of **data1_A** are multiplied by the corresponding elements of **data1_B** and the product of each multiplication is put in the corresponding rows of the **data1_C** data set.

The resultant worksheet displays as follows:



	A[X]	B[Y]	C[Y]
1	1	7	7
2	2	8	16
3	3	9	27
4	4		–
5	5		–
6	6		–

An example of a vector calculation involving a scaler follows. You can also execute this LabTalk script in the Script window:

```
data1_b = 3 * data1_a
```
(press ENTER)

In this example, every element of the **data1_a** data set is multiplied by 3. The results are put into the corresponding rows of the **data1_b** data set.

If you were to instead execute the following script (assuming `newData` does not previously exist) in the Script window:

```
newData = 3 * data1_A
```
(press ENTER)

then a temporary data set called `newData` is created and assigned the result of the vector operation. Elements of the temporary data set can be accessed in the same way you would access an element of a data set contained in a worksheet.

Calculations Using Interpolation

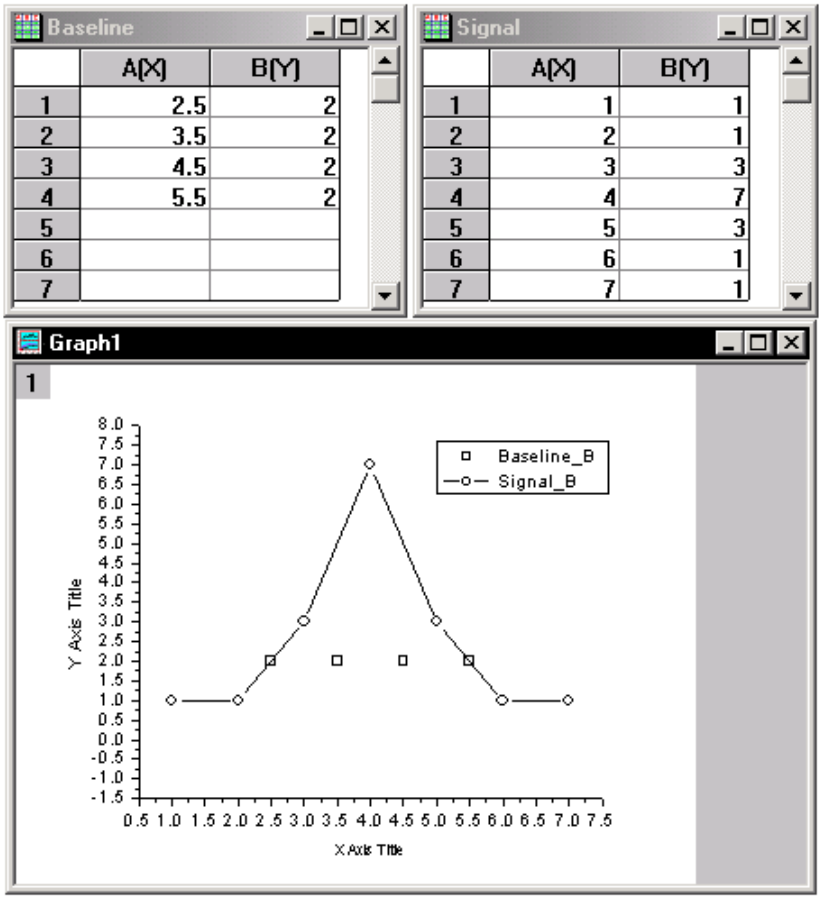
When you use one of the following notations to perform vector operations:

```
data set = scaler operator data set
```

or

```
data set = data set operator data set
```

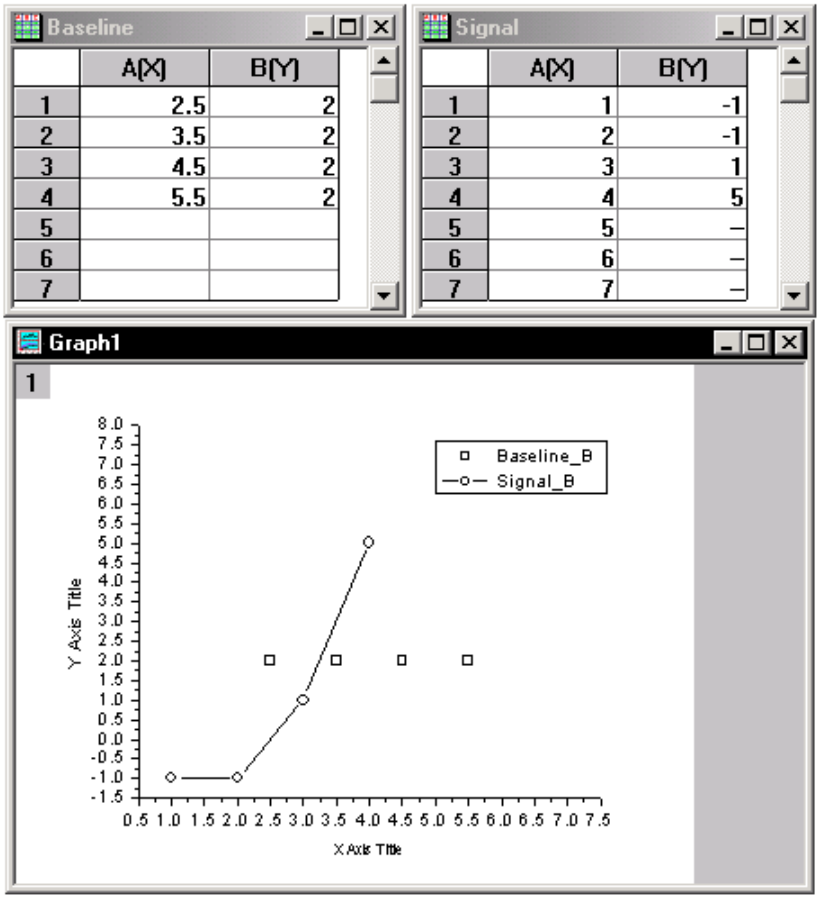
row by row calculations are performed even if the data sets have different numbers of elements. However, there may be times when you want to use linear interpolation when performing calculations on data sets of different sizes. As an example, consider the following two dependent data sets, each with its own set of independent data:



If you were to subtract **baseline_b** from **signal_b** by entering the following script in the Script window:

```
signal_b -= baseline_b
```
(press ENTER)

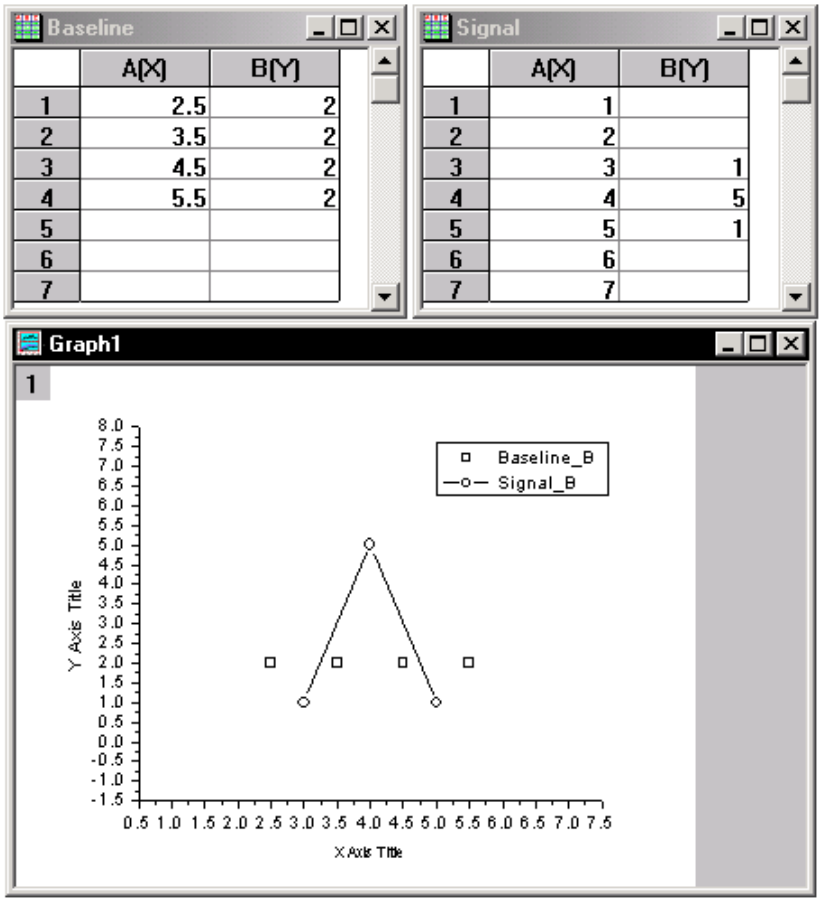
then the subtraction would be performed row by row (with no linear interpolation) as follows:



However, if you use a *`data set operator data set`* assignment-operation statement (starting with the initial worksheet values) in the Script window:

`signal_b - baseline_b` (press ENTER)

then linear interpolation of the second data set (**baseline_b**) within the domain of the first data set (**signal_b**) is used to carry out the subtraction. The results are stored in the first data set (**signal_b**). The following figure shows the results of this calculation, when starting with the initial worksheet values.



In this example, interpolation was used to find the Y values of **baseline_b** using the X values given by **signal_a** *within the X domain* of **baseline_b**.

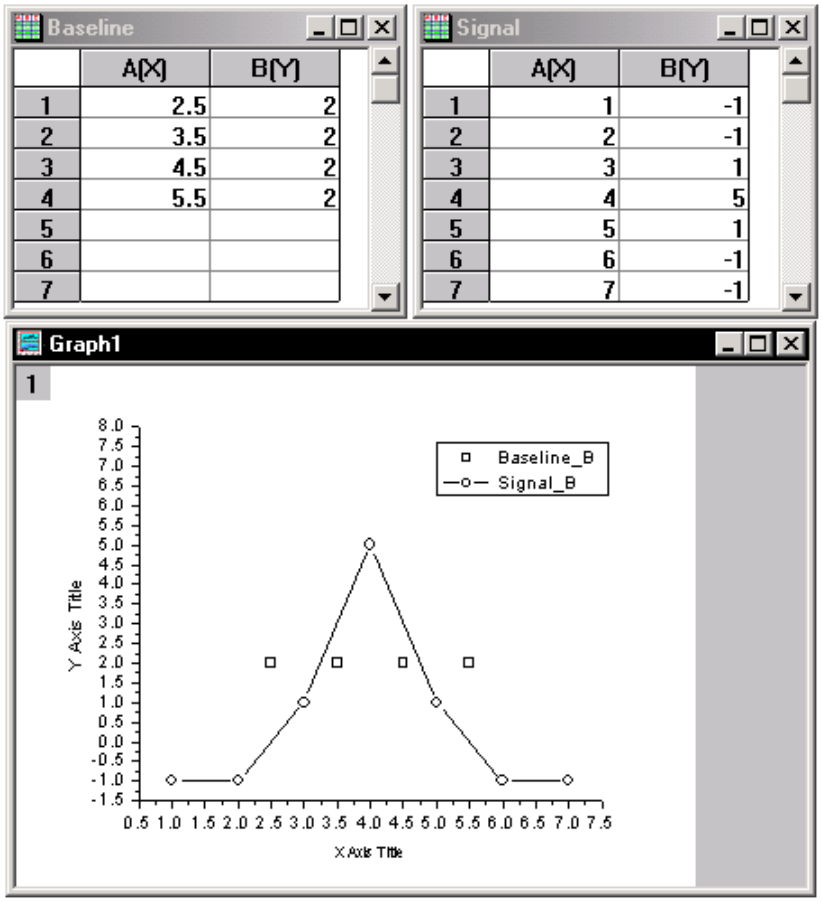
In order to have Origin perform the interpolation *outside the domain* of the second data set (in this case, **baseline_b**) for all X values of the first data set, you need to specify an option switch (`-O`). The syntax using this option switch is:

```
data set operator-O data set
```

Thus, in this example, you would enter the following script in the Script window:

```
signal_b --O baseline_b
```
(press ENTER)

The following figure shows the results of this calculation, when starting with the initial worksheet values.



In this example, the option switch caused linear interpolation of **baseline_b** to be used within and outside of its domain in order to perform the calculation for the entire domain of **signal_b**.

In addition to using the subtraction operator with the `-O` option, you can use the addition (+), multiplication (*), division (/), and exponentiate (^) operators with `-O`.

Note: When data is unsorted or there are duplicate Y values for any given X value, there will be inconsistencies in the linear interpolation results.

For more information on LabTalk, see the LabTalk Language Reference section of the Programming Help file (**Help:Programming**).

Subtracting Reference Data

To subtract one data plot from another, select **Analysis:Subtract:Reference Data**. This menu command is only available when a graph is active.

This menu command opens the Math on\between Data set dialog box with the (-) operator selected in the Operator text box. Select and move the desired data sets into the Y1 and Y2 text boxes using the => buttons. If the X values of the two data sets do not match, interpolation is automatically performed on the reference data. Because this menu command changes the Y data set in the worksheet, it affects all plot instances of the data set.

Subtracting a Straight Line

To define a straight line and then subtract that line from the active data plot, select **Analysis:Subtract Straight Line**. This menu command is only available when a graph is active.

When this menu command is selected, the cursor changes to a cross-hair symbol (the Screen Reader tool). The Data Display tool opens if it is not already open. Click in the graph window to view the XY coordinates of the selected location in the Data Display tool. Double-click at the desired location to set the endpoints of the line. Origin subtracts the line you have drawn from the active data plot and plots the result. Because this menu command changes the Y data set in the worksheet, it affects all plot instances of the data set.

Translating Vertically or Horizontally

Vertical Translation

To move the entire active data plot vertically along the Y axis, select **Analysis:Translate:Vertical**. This menu command is only available when a graph is active.

This menu command opens the Data Display tool if it is not already open. Double-click on the data plot (or click and press ENTER) to choose the initial data point. The XY coordinates of the data point display in the Data Display tool. Double-click again at any point on the page. Origin calculates the difference in Y between the two selected Y-coordinates and adds this amount to the entire active data plot. Because this menu command changes the Y data set in the worksheet, it affects all plot instances of the data set.

Horizontal Translation

To move the entire active data plot horizontally along the X axis, select **Analysis:Translate:Horizontal**. This menu command is only available when a graph is active.

This menu command opens the Data Display tool if it is not already open. Double-click on the data plot (or click and press ENTER) to choose the initial data point. The XY coordinates of the data point display in the Data Display tool. Double-click again at any point on the page. Origin calculates the difference in X between the two selected X-coordinates and adds this amount to the entire active data plot. Because this menu command changes the X data set in the worksheet, it affects all plot instances of the data set.

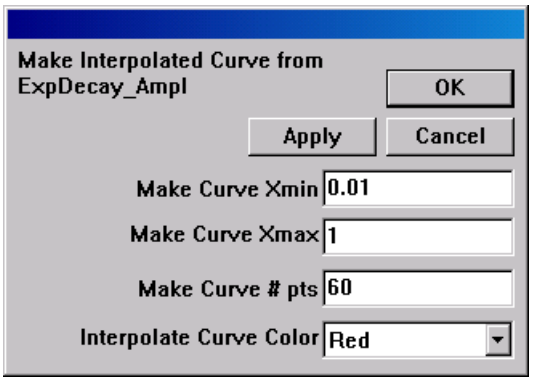
Averaging Multiple Curves

To calculate the average Y value for all data plots in the active layer at each X coordinate of the active data plot, select **Analysis:Average Multiple Curves**. The result is saved as a new data set (in a new Average# hidden worksheet) and plotted in the active layer. The new data set is named **average#_mean*N*cuv** where

is the average curve number in the project (starting from 1) and *N* is the number of data plots used to calculate the average. If necessary, Origin will interpolate or extrapolate data plots when calculating the average Y value.

Interpolating and Extrapolating

To interpolate or extrapolate between the data points of the active data plot, select **Analysis:Inter/Extrapolate**. This menu command is only available when a graph is active. The menu command opens the Make Interpolated Curve from *Dataset* dialog box.



Type the minimum and maximum values to set the length of the interpolated or extrapolated data plot in the Make Curve Xmin and Make Curve Xmax text boxes. You are not limited to the original length.

Specify the number of points to draw in the interpolated or extrapolated data plot in the Make Curve # Points text box.

Click Apply to view the interpolated or extrapolated data plot based on the current settings including the specified Interpolate Curve Color. The dialog box remains open and is available for further changes. Each time you change a setting and click Apply, the results update in the graph. The results are not finalized until you click OK

After clicking OK, the result is saved as a new data set (in a new InterExtrap# hidden worksheet) and plotted in the active layer. The new data set is named **InterExtrap#_*wkscol*** where # is the inter/extrapolated curve number in the project (starting from 1) and *wkscol* is the concatenated worksheet and column name of the original data set.

Note that the interpolation or extrapolation is dependent on the type of line connecting the data points in the original data plot.

To learn how to perform calculations on data sets (row by row or using interpolation) using LabTalk, see "Performing Calculations on Data Sets Using LabTalk" on page 440.

Differentiating

To calculate the derivative of the active data plot, select **Analysis:Calculus:Differentiate**. Origin calculates the derivative and adds this data to a new Derivative# (hidden) worksheet. Origin then opens an

instance of the DERIV.OTP graph template and plots the derivative into this window. Any subsequent derivative operation is plotted into this Deriv window.

The derivative is taken by averaging the slopes of two adjacent data points as follows:

$$\frac{1}{2}\left(\frac{y_{i+1}-y_i}{x_{i+1}-x_i}+\frac{y_i-y_{i-1}}{x_i-x_{i-1}}\right)$$

Differentiating using Savitzky-Golay Smoothing

To find the first or second order derivatives for the active data plot, select **Analysis:Calculus:Diff/Smooth**. This menu command opens the Smoothing dialog box. Specify the Polynomial Order (from 1-9), the Points to the Left, and the Points to the Right, and click OK. If you select a Polynomial Order greater than 1, the Derivatives On *Dataset* dialog box opens. Specify the derivative order and click OK. An instance of the DERIV.OTP graph template then opens. The derivative is plotted into this window. The derivative data is saved in a new Smoothed# hidden worksheet, where # is the derivative curve number in the project (starting from 1).

The **Diff/Smooth** menu command uses the Savitzky-Golay smoothing method that performs a local polynomial regression around each point. Enough information is thus provided to calculate the derivatives of the data set up to an order equal to the order of the polynomial used (in Origin's case, this order is 2). This "side-effect" (of the Savitzky-Golay smoothing) is used by the **Diff/Smooth** menu command to find either the first or second derivative of the active data plot. The derivatives thus calculated are the coefficients of the local polynomials that approximate the data plot.

Integrating

To numerically integrate the active data plot from a baseline of zero using the trapezoidal rule, select **Analysis:Calculus:Integrate**. The resulting area, peak height (maximum deflection from the X axis), peak position, and peak width are displayed in the Results Log.

The integrated result is stored in the data set **_integ_area**. Because this is a temporary data set, it must be copied to a non-temporary data set in order to be manipulated. (All temporary data sets are deleted when the project is printed or saved.) To copy the temporary data set to a non-temporary data set, and then view the new data set in the worksheet, type the following script in the Script window (replace ***MyNewWks_b*** with the desired worksheet and column name):

copy -x _integ_area *MyNewWks_b* (ENTER)

edit *MyNewWks_b* (ENTER)

Note 1: By convention, reversing the limits of integration changes the sign of the integral. When integrating over an interval [a,b], it is presumed that a<b. Therefore, when a>b (x values are in descending order), integration yields a negative area.

Note 2: Origin also supports 3D integration. To compute the volume beneath the surface defined by the matrix, select **Matrix:Integrate** when a matrix is active. Origin performs a double integral over X and Y to compute the volume and reports the value in the Script window.

Statistics

Descriptive Statistics

Statistics on Columns

To perform statistics on worksheet data, highlight the desired columns or range of cells, and select **Statistics:Descriptive Statistics:Statistics on Columns**. This menu command opens a new worksheet that displays the mean, standard deviation, standard error of the mean, minimum, maximum, range, sum, and number of points for each of the highlighted columns (or ranges) in the active worksheet. The standard deviation and standard error of the mean columns are automatically set as error bars to direct plotting.

Note: Columns that are set to Disregard from the Plot Designation drop-down list of the Worksheet Column Format dialog box as well as columns that are set to Text from the Display drop-down list of this dialog box are not used in the calculation.

The new worksheet contains a Recalculate button to recalculate the statistics data if column values change. Additionally, an Advanced Statistics check box is provided. Select this check box and then click Recalculate to find the median, 25th and 75th percentiles (lower and upper quartiles), user-defined percentile (specified in the Percentile text box), Inter Quartile Range, variance, 95% confidence limits on the mean, and Kurtosis.

The standard deviation (SD) is calculated as follows:

$$SD = \sqrt{Var}$$

where $Var = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$, n is the sample size, and $\overline{X}$ is the mean.

The standard error of the mean (SEM) is calculated as follows:

$$SEM = \frac{SD}{\sqrt{n}}$$

Statistics on Rows

To perform statistics on rows of data, highlight the desired cells (including columns of data) and select **Statistics:Descriptive Statistics:Statistics on Rows**. This menu command creates a new worksheet

displaying the row number, mean, standard deviation, standard error of the mean, minimum, maximum, range, and number of points for the selected data.

Note: Columns that are set to Disregard from the Plot Designation drop-down list of the Worksheet Column Format dialog box as well as columns that are set to Text from the Display drop-down list of this dialog box are not used in the calculation.)

Frequency Count

To determine a frequency count for a column or range of values, select **Statistics:Descriptive Statistics:Frequency Count**. This menu command opens the Count *Dataset* dialog box in which you specify the minimum and maximum values and the step increment. From this information, Origin creates a set of bins, beginning at the minimum value and based on the step size. Origin then searches each bin, and records the number of times a value is encountered that falls within each bin. If a value falls *on the upper edge of the bin, it is included in the next higher bin*.

Origin then creates a worksheet with four columns. The first column displays the mid-point value in each bin, the second column displays the number of "hits" observed for each bin, the third column displays the end point of each bin, and the fourth column displays the cumulative count. The Results Log displays the mean, standard deviation, size, and median value for the selected data. The median value is calculated by sorting the data.

Normality Test (Shapiro-Wilk)

The Shapiro-Wilk Normality Test is used to determine whether or not a random sample of values (X_i for i = 1 to n) follows a normal distribution. The Normality Test is useful because other statistical tests (such as the t-Test) assume that data is sampled from a normally distributed population. A W statistic and a P value are computed, from which a statistical decision can be made by comparison with a level of significance. The W statistic is defined as:

$$W = \frac{\left(\sum_{i=1}^{n} A_i X_i\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and A_i is a weighting factor. The W value reported by Origin is computed using the NAG function **nag_shapiro_wilk_test (g01ddc)** that implements the Applied Statistics Algorithm AS 181 described in Royston (1982). The function supports sample sizes of $3 \le n \le 2000$. See the NAG documentation for more detailed information. (When you install Origin, you are provided the option to install the NAG PDF files which document the NAG functions. If you clicked Yes to install these files, a \NAG PDFs folder is created with a subfolder for each library. If you did not install the PDFs, they remain accessible on your installation CD. Furthermore, you can install them at a later date by running the Origin "Add or Remove Files" program.)

Performing the Shapiro-Wilk Normality Test

To activate the Shapiro-Wilk Normality Test, highlight one or more columns of non-text data in a worksheet and then select **Statistics:Descriptive Statistics:Normality Test (Shapiro-Wilk)**. The test is performed and the data set name, sample size, W statistic, P value, and decision rule (based on a default significance level of 0.05) is output for each selected data set.

To specify a significance level other than 0.05, open the Script window (**Window:Script Window**) and assign a value between 0 and 100 to the LabTalk variable `ONormTestSL` prior to executing the Normality Test menu command. For example, entering:

```
ONormTestSL=10   (Press ENTER)
```

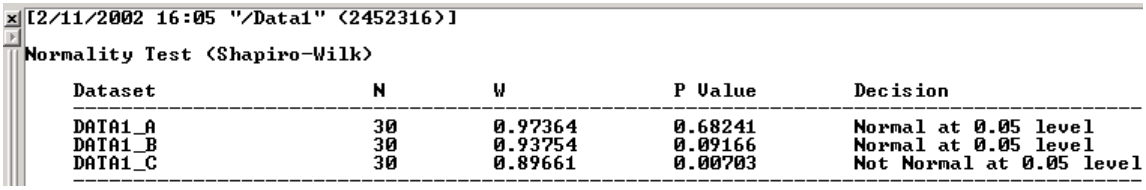
will set the Normality Test significance level to 0.1.

Example: Performing a Shapiro-Wilk Normality Test

1) To perform a Shapiro-Wilk Normality Test, open the Origin sample project file NORMALITY TEST.OPJ located in the Origin \SAMPLES\ANALYSIS\STATISTICS subfolder.

2) Highlight all three columns of data in the Data1 worksheet and select **Statistics:Descriptive Statistics:Normality Test (Shapiro-Wilk)**.

The test is performed and the following results are output in the Results Log.



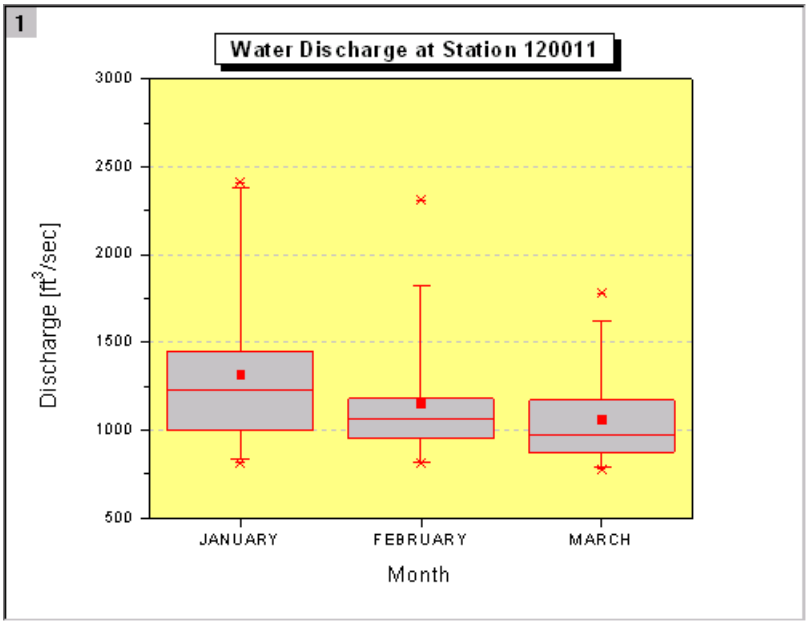
[2/11/2002 16:05 "/Data1" (2452316)]

Normality Test (Shapiro-Wilk)

Dataset	N	W	P Value	Decision
DATA1_A	30	0.97364	0.68241	Normal at 0.05 level
DATA1_B	30	0.93754	0.09166	Normal at 0.05 level
DATA1_C	30	0.89661	0.00703	Not Normal at 0.05 level

Creating Box Charts

To create a box chart, highlight one or more Y worksheet columns (or a range from one or more Y columns) and select **Plot:Statistical Graphs:Box Chart**.



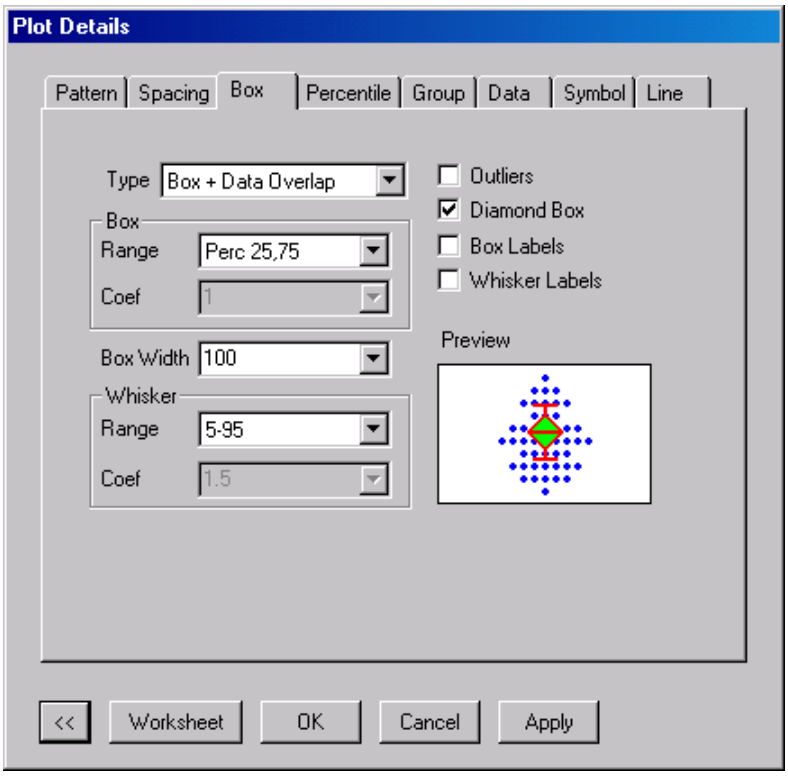
Each Y column of data is represented as a separate box. The column names or labels supply the X axis tick labels. By default, the box is determined by the 25th and 75th percentiles. The whiskers are determined by the 5th and 95th percentiles.

Note: To learn more about box charts, review the BOX CHART.OPJ project located in your Origin \SAMPLES\GRAPHING\STATISTICAL GRAPHS folder.

Customizing the Box Chart

To customize the box chart display, double-click on the box chart, or right-click and select **Plot Details**. Both actions open the Plot Details dialog box with the box chart data plot icon active on the left side of the dialog box. The box chart controls are available on tabs on the right side of the dialog box.

The Box Tab on the Plot Details Dialog Box



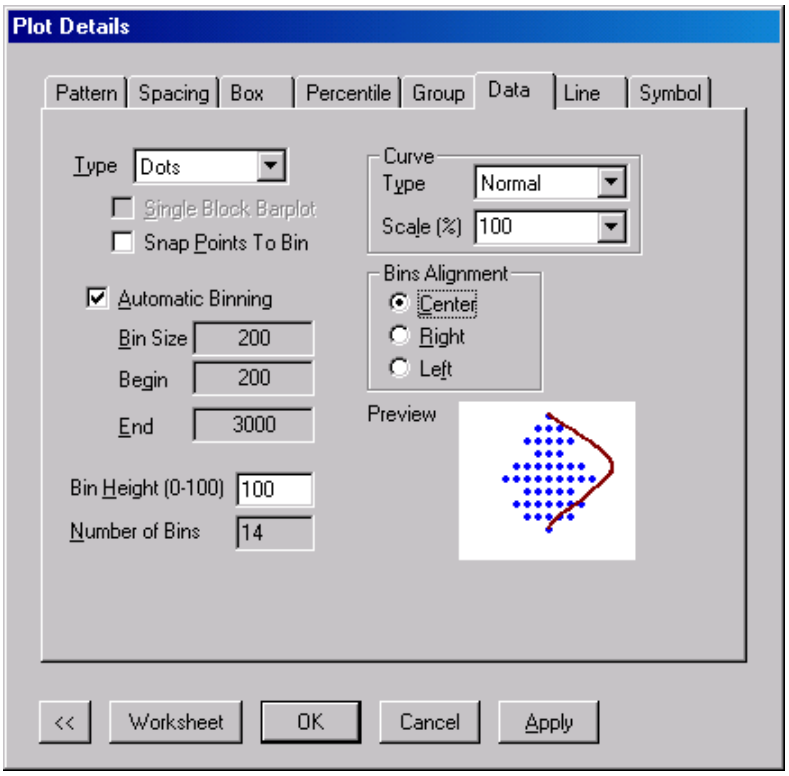
By default, the box chart displays only the box - not the binned data. You can, however, display the box and binned data, or only the binned data. The binned data is saved in a Bin*n* worksheet. This worksheet contains the bin X values, counts, cumulative sum, and percentages. (To access the Bin*n* worksheet, right-click on the box chart and select **Go to Bin Worksheet** from the shortcut menu.)

To display the binned data with the box chart, or only the binned data, select the desired display from the Type drop-down list When you select a display that includes the binned data, the Plot Details Data tab becomes available for editing the bins and overlaying a distribution curve.

The Box group and the Whiskers group control the display of the box and whiskers. Select the desired percentiles or other values to represent the box or whiskers.

Outliers, diamond box, and box and whisker labels are activated by selecting the associated check boxes on the Box tab.

The Data Tab on the Plot Details Dialog Box



The Data tab is available when you select a box chart display that includes the binned data from the Type drop-down list on the Box tab. Edit this tab to control the display of the binned data.

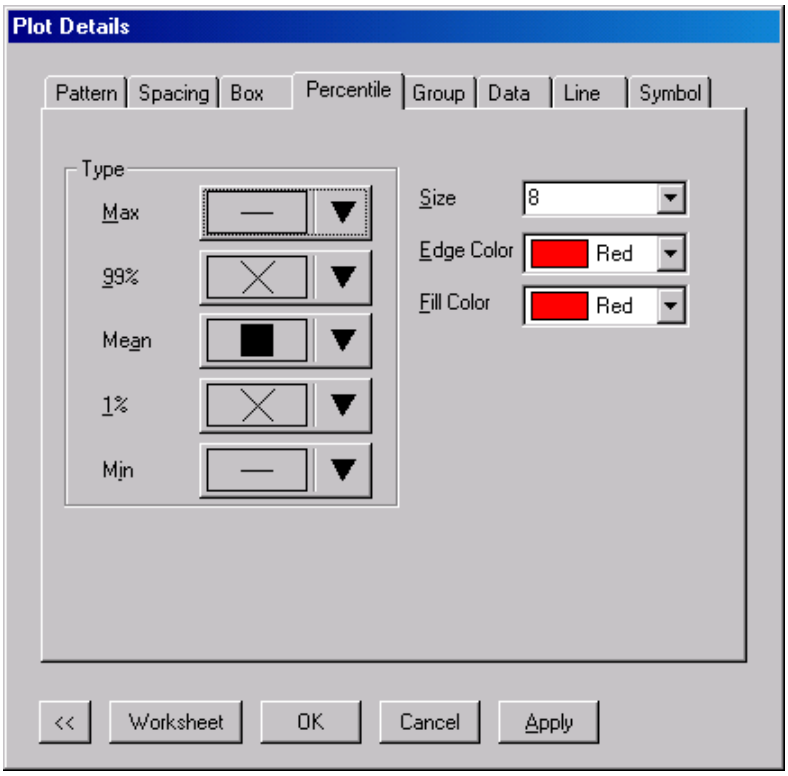
Select Dots, Bars, or Dots and Bars from the Type drop-down list. Dots shows the individual binned data points and bars shows a histogram-like representation of the data, with bars of the specified bin size.

Select the Single Block Bar Plot check box to remove the lines between bins when Bars is selected from the Type drop-down list. Select the Snap Points to Bin check box to align the binned data points horizontally in the bins.

Edit the bin limits by clearing the Automatic Binning check box and editing the associated text boxes.

Overlay a distribution curve on the binned data by selecting Normal, Lognormal, Poisson, Exponential, Laplace, or Lorentz from the Type drop-down list. The Preview window displays your selection. Curves selected from the Type drop-down list are not "fitted" to your data. Instead, Origin checks where the mean of your data is, then overlays the curve, with its mean in the same place. If it is a two parameter curve, then Origin accounts for the standard deviation of your data in its curve also. The Scale combination box in the Curve group tells Origin how much of the box chart's drawing space in the X direction you want the curve to take up. That is to say, if you have 100 selected from this drop-down list, then all the drawing space in the X direction will be used (the curve will be wider). If you have 50 selected, then the curve will only take up half of the allocated drawing space (the curve will be narrower).

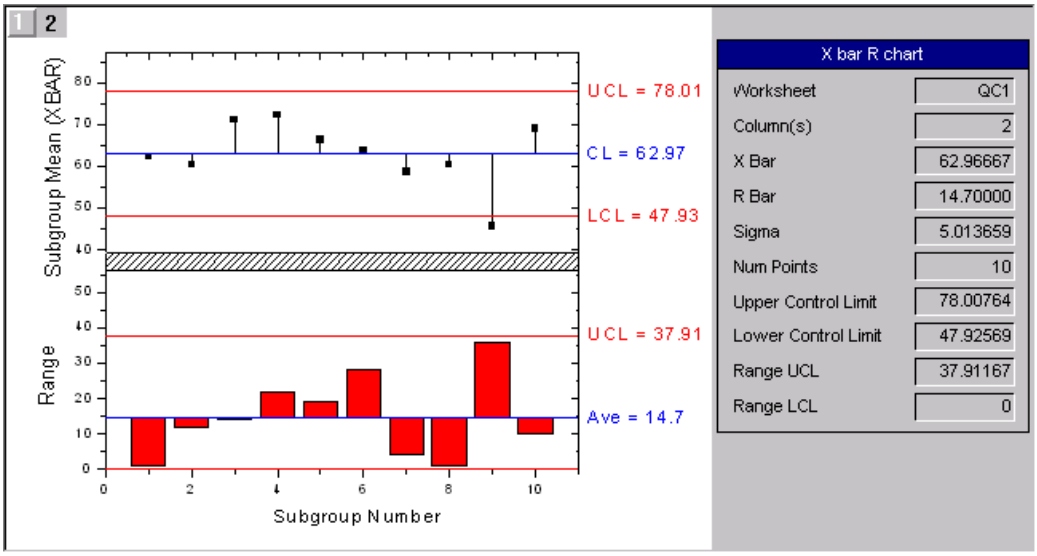
The Percentile Tab on the Plot Details Dialog Box



The Percentile tab provides controls for customizing the symbols for various percentiles in the box chart. Click the down arrow next to the symbol you want to change. Select the desired symbol from the symbol gallery that opens.

Creating QC Charts

QC charts are used to study the fluctuation of data in a continuous process. To create a QC chart, highlight at least one column of values (or a range from at least one column) and select **Plot:Statistical Graphs:QC (X bar R) Chart**. Origin opens the X bar R Chart dialog box in which you specify the subgroup size for the selected data set. Origin then creates a worksheet and graph window displaying two layers.



The worksheet contains the mean, range, and standard deviation for each subgroup in the selected data set.

The upper layer of the QC chart is the X bar graph. This layer displays the mean value for each of the subgroups as a scatter graph with drop lines to the average of the mean for each group (X bar). This layer also displays two limit lines that are positioned *Num Sigma* standard deviations away from the X bar. *Num Sigma* is defined in the QC1 worksheet window and is three by default.

The lower layer of the QC chart is the R chart. This layer displays the range for each of the subgroups as a column graph. The range for each of the subgroups is plotted from the average line (or R bar). This line represents the groups average range, or the average of the mean range within each subgroup. This layer also displays two limit lines - the UCL and LCL lines. For information on the calculations for the UCL and LCL lines, see "Statistical Quality Control" 5th edition (1980) by Grant, E.R. and Leavenworth, R.S., McGraw Hill, pg 81.

To create a new QC chart based on the original worksheet, edit the desired text boxes in the QC1 worksheet and then click the Make QC Chart button. To display multiple QC charts, type a new window name in the Plot text box.

Note: To learn more about QC charts, review the QC CHART.OPJ project located in your Origin \SAMPLES\GRAPHING\STATISTICAL GRAPHS folder.

Creating a QC Chart from Multiple Data Sets

To create a QC chart from more than one data set in the worksheet, highlight the desired worksheet columns and select **Plot:Statistical Charts:QC (X bar R) Chart**. An Attention dialog box informs you that a group is defined for each row, which extends over the selected column range.

Creating Histograms

A histogram displays the number of times a data value in a selected column fell within a specified bin. To create a histogram, highlight one or more Y worksheet columns (or a range from one or more Y columns) and select **Plot:Statistical Graphs:Histogram**. Origin automatically calculates the bin size and creates a new graph from the HISTGM.OTP template. The binned data is saved in a Bin*n* worksheet. This worksheet contains the bin X values, counts, cumulative sum, and percentages. (To access the Bin*n* worksheet, right-click on the histogram and select **Go to Bin Worksheet** from the shortcut menu.)

Note: The **Histogram** menu command plots each selected data set in the same layer. The **Stacked Histograms** menu command plots each selected data set in its own layer (using the same bin limit for each layer).

To learn more about histograms, review the HISTOGRAM.OPJ project located in your Origin \SAMPLES\GRAPHING\STATISTICAL GRAPHS folder.

Customizing the Histogram

To customize the histogram display, double-click on the histogram, or right-click and select **Plot Details**. Both actions open the Plot Details dialog box with the histogram data plot icon active on the left side of the dialog box. The histogram controls are available on tabs on the right side of the dialog box.

Histogram with Probabilities

To create a histogram with probabilities, highlight a single worksheet column (or a range from a worksheet column) and select **Plot:Statistical Graphs:Histogram + Probabilities**. This menu command is nearly identical to the **Histogram** menu command, except that the template invoked by this command (HISTCUMU.OTP) also displays a cumulative sum of the data (layer 2).

Additionally, the **Histogram + Probabilities** menu command records the statistical results to the Results Log. The statistical results include the mean, the standard deviation, the maximum and minimum values, and the total number of values.

One Sample t-Tests

A one sample t-Test can be employed to test whether or not the true mean of a population, μ, is equal to or different than a specified test mean, μ_0. The one sample t-Test is performed on a sample data set that is randomly drawn from a population of scores that is assumed to follow a normal distribution.

The one sample t-Test can either be a one- or two-tailed test depending on the nature of the variable under study and the objectives of the experimenter. The experimenter develops a null hypothesis (H_0) that is the logical counterpart, mutually exclusive and exhaustive, to an alternative hypothesis (H_1) which the experimenter is attempting to evaluate. Depending on the outcome of the test, the null hypothesis is either rejected or retained. Rejection of the null hypothesis logically leads to acceptance of the alternative hypothesis and retention of the null hypothesis leads to an inability to accept the alternative hypothesis. Two-tailed hypotheses take the form:

$H_0: \mu = \mu_0$ and $H_1: \mu \neq \mu_0$

while one-tailed hypotheses take the form:

$H_0: \mu \leq \mu_0$ and $H_1: \mu > \mu_0$ or $H_0: \mu \geq \mu_0$ and $H_1: \mu < \mu_0$

depending on the direction of the test.

The test statistic t is used to evaluate the null hypothesis which then allows the indirect evaluation of the alternative hypothesis. The computed test statistic, $t = \frac{\left(\overline{X} - \mu_0\right)}{SD/\sqrt{n}}$, has a Student's t-distribution with $\nu = n - 1$ degrees of freedom, where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $SD = +\sqrt{\sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{n-1}}$. We reject the null hypothesis H_0 (and thereby accept the alternative hypothesis H_1) when t is extreme enough to cause the observed significance to be less than some pre-specified significance level. We retain the null hypothesis (and thereby fail to accept the alternative hypothesis) when the observed significance is greater than the pre-specified significance level. The observed significance (or P value) is the probability of obtaining a test statistic as extreme or more extreme than t due to chance factors alone. The pre-specified significance level is called alpha (α) or just significance level.

In the context of a one sample t-Test, a confidence interval consists of a lower limit and an upper limit that, with some specified degree of confidence, contains the true mean of the population. The degree of confidence (or probability) that a given confidence interval contains the true mean of a population is called the confidence level. The lower and upper limits of a confidence interval are $\overline{X} \pm \frac{t_{1-\alpha/2}(\nu)SD}{\sqrt{n}}$ where $\overline{X}$ and SD are defined as above. The factor $t_{1-\alpha/2}(\nu)$ is a critical value of t from the Student's t distribution indexed at the $1 - \alpha/2$ level and by $\nu = n - 1$ degrees of freedom. Here alpha (α) is specified by the confidence level where α = 1 - (*confidence level*/100).

The power of a one sample t-Test is a measurement of its sensitivity. In terms of the null and alternative hypotheses, power is the probability that the test statistic t will be extreme enough to allow the rejection of the null hypothesis when it should in fact be rejected (i.e. given the null hypothesis is not true). For each of the three different null hypotheses, power is mathematically defined as:

$\mu = \mu_0: \; P[t \leq -t_{1-\alpha/2}(\nu) \mid \mu] + P[t \geq t_{1-\alpha/2}(\nu) \mid \mu]$

$\mu \leq \mu_0: \; P[t \geq t_{1-\alpha}(\nu) \mid \mu]$

$\mu \geq \mu_0: \; P[t \leq -t_{1-\alpha}(\nu) \mid \mu]$

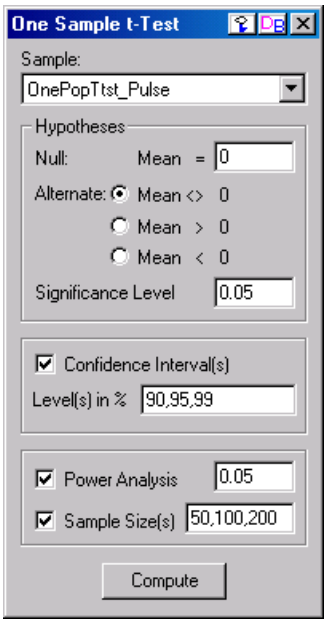
where the computation for the test statistic t is given above, the factors $t_{1-\alpha/2}(\nu)$ (or $t_{1-\alpha}(\nu)$) are critical values of t from the Student's t distribution indexed at the $1 - \alpha/2$ (or $1 - \alpha$) level and by $\nu = n - 1$

degrees of freedom, and μ is the true population mean. The computation for hypothetical power is the same as for actual power except that t and ν are recomputed using hypothetical sample sizes instead of the actual sample size.

Performing a One Sample t-Test

To perform a one sample t-Test, highlight the desired worksheet column and then select **Statistics:Hypothesis Testing:One Sample t-Test**. The menu command opens the One Sample t-Test dialog box in which you specify the Sample data set as well as all other test parameters described below. After clicking the Compute button, the test results are output to the Results Log. The results include the sample data set's name, mean ($\overline{X}$), standard deviation (SD), standard error (SE), and sample size (n). The test statistic (t), degrees of freedom (DF or ν), observed significance (P value), the null (H_0) and alternative (H_1) hypotheses, and the decision rule of the test are output as well. If the Confidence Interval(s) and Power Analysis check boxes are enabled, then all specified confidence intervals and the actual power of the experiment are output. Checking the Sample Size(s) check box causes a hypothetical power for each specified sample size to also be output.

Operating the One Sample t-Test Controls



1) The Sample Combination Box

Select the sample data set on which to perform a one sample t-Test. The sample data is assumed to have been drawn from a population that follows a normal distribution.

2) The Hypotheses Group

Enter the test mean, μ_0 , in the Null Hypothesis edit box just to the right of the text "Mean = ". The value of the test mean displayed in each of the three Alternate Hypotheses updates accordingly.

Select the appropriate radio button for Alternate Hypothesis indicating whether a two-tailed or one-tailed t-Test is to be performed. The direction of the one-tailed t-Test is also specified by the radio button selection. The (in)equality of the Null Hypothesis updates accordingly.

Enter a decimal value greater than 0 and less than 1 for alpha (α) in the Significance Level edit box. The Power Analysis edit box in the Power Analysis group is initialized to the same value of alpha (α) and, if unique, a complementary ([1-α]$*$100) confidence level is added to the Level(s) in % edit box in the Confidence Interval(s) group.

3) The Confidence Interval(s) Group

Select or clear the Confidence Interval(s) check box to determine whether or not confidence intervals are computed. The Level(s) in % edit box is enabled when the Confidence Interval(s) check box is selected.

Enter a comma separated list of confidence levels greater than 0 and less than 100 in the Level(s) in % edit box.

4) The Power Analysis Group

Select or clear the Power Analysis check box to determine whether or not the "actual" power of the t-Test is computed. The Power Analysis edit box (located to the right of its check box) and the Sample Size(s) check box are enabled when the Power Analysis check box is selected.

Enter a decimal value greater than 0 and less than 1 for alpha (α) in the Power Analysis edit box. By default, the Power Analysis edit box is initialized to the same value of alpha (α) that is entered in the Significance Level edit box but it can be modified if desired. The "actual" power computation uses the sample size of the selected sample data and the alpha (α) specified in the Power Analysis edit box.

Select or clear the Sample Size(s) check box to determine whether or not "hypothetical" powers for the t-Test are computed. The Sample Size(s) edit box (located to the right of its check box) is enabled when the Sample Size(s) check box is selected.

Enter a comma separated list of hypothetical sample sizes (positive Integers only) in the Sample Size(s) edit box. A "hypothetical" power is computed for each hypothetical sample size entered. The hypothetical power computation uses the alpha (α) specified in the Power Analysis edit box.

5) The Compute Button

Click the Compute button to perform the indicated computations on the selected sample data set. Control settings (including the Sample) can be changed and the Compute button can be clicked as many times as desired without closing the dialog box. All results are output to the Results Log.

Example: Performing a One Sample t-Test

1) To perform a one sample t-Test, open the Origin sample project file ONE SAMPLE T-TEST.OPJ located in the Origin \SAMPLES\ANALYSIS\STATISTICS subfolder. This data file contains 130 observations of normal body temperature (°F), gender (1=male, 2=female), and heart rate (beats/minute).

2) Highlight the leftmost column of values in the OnePopTtst worksheet and select **Statistics:Hypothesis Testing:One Sample t-Test**.

3) Type **98.6** for μ_0 in the Null Hypothesis edit box and check the Confidence Interval(s), Power Analysis, and Sample Size(s) check boxes.

4) Accept all other default values and click the Compute button.

One Sample t-Test Results

The observed significance (P value = 2.41063E-7) gives the probability of obtaining a test statistic as extreme or more extreme than t by chance alone if the sampled data were in fact drawn from a population having a mean (μ) equal to 98.6°F. Thus, with a two-tailed significance level (or alpha) set at $\alpha = 0.05$, we reject the null hypothesis and accept the alternative hypothesis. We conclude that the true population mean is not 98.6°F.

In addition, the 99% confidence interval tells us that we can be 99% sure that the limits 98.08111 (lower) and 98.41735 (upper) contain the true population mean. Power analysis (actual power = 0.99965) leads us to conclude that the experiment is very sensitive. Given that the true population mean is not not 98.6°F, it is highly probable that this experiment would correctly reject the null hypothesis (as it did). We might even be able to reduce the sample size (and cost) of similar experiments to 100 (hypothetical power = 0.99693) without loosing much sensitivity. Increasing the sample size to 200 would cause power to approach 1 (with 5 decimal digits of precision) but we would have to incur additional expense and gain little sensitivity.

Two Sample t-Tests

A two sample t-Test can be employed to test whether or not two population means, μ_1 and μ_2, are equal (i.e. whether or not their difference is 0, $\mu_1 - \mu_2 = d_0 = 0$). The two sample t-Test is performed on two sample data sets that are assumed to have been drawn from populations that follow a normal distribution with constant variance.

The two sample t-Test can either be a one- or two-tailed test depending on the nature of the variable under study and the objectives of the experimenter. The experimenter develops a null hypothesis (H_0) which is the logical counterpart, mutually exclusive and exhaustive, to an alternative hypothesis (H_1) that the experimenter is attempting to evaluate. Depending on the outcome of the test, the null hypothesis is either rejected or retained. Rejection of the null hypothesis logically leads to acceptance of the alternative hypothesis and retention of the null hypothesis leads to an inability to accept the alternative hypothesis. Two-tailed hypotheses take the form:

$H_0: \mu_1 - \mu_2 = d_0$ and $H_1: \mu_1 - \mu_2 \neq d_0$

while one-tailed hypotheses take the form:

$H_0: \mu_1 - \mu_2 \leq d_0$ and $H_1: \mu_1 - \mu_2 > d_0$

or

$H_0: \mu_1 - \mu_2 \geq d_0$ and $H_1: \mu_1 - \mu_2 < d_0$

depending on the direction of the test.

The test statistic t is used to evaluate the null hypothesis which then allows the indirect evaluation of the alternative hypothesis. We reject the null hypothesis H_0 (and thereby accept the alternative hypothesis H_1) when t is extreme enough to cause the observed significance to be less than some pre-specified significance level. We retain the null hypothesis (and thereby fail to accept the alternative hypothesis) when the observed significance is greater than the pre-specified significance level. The observed significance (or P value) is the probability of obtaining a test statistic as extreme or more extreme than t due to chance factors alone. The pre-specified significance level is called alpha (α) or just significance level.

In the context of a two sample t-Test, a confidence interval consists of a lower limit and an upper limit that, with some specified degree of confidence, contains the difference of the true population means. The degree of confidence (or probability) that a given confidence interval contains the difference of the true population means is called the confidence level.

Note: See the separate discussions for independent and paired t-tests below for their respective mathematical definitions of the test statistic t and confidence intervals.

The power of a two sample t-Test is a measurement of its sensitivity. In terms of the null and alternative hypotheses, power is the probability that the test statistic t will be extreme enough to allow the rejection of the null hypothesis when it should in fact be rejected (i.e. given the null hypothesis is not true). For each of the three different null hypotheses, power is mathematically defined as:

$$\mu_1 - \mu_2 = d_0: \; P[t \leq -t_{1-\alpha/2}(\nu) \mid \mu_1 - \mu_2] + P[t \geq t_{1-\alpha/2}(\nu) \mid \mu_1 - \mu_2]$$

$$\mu_1 - \mu_2 \leq d_0: \; P[t \geq t_{1-\alpha}(\nu) \mid \mu_1 - \mu_2]$$

$$\mu_1 - \mu_2 \geq d_0: \; P[t \leq -t_{1-\alpha}(\nu) \mid \mu_1 - \mu_2]$$

where the computation for the test statistic t is given below, the factors $t_{1-\alpha/2}(\nu)$ (or $t_{1-\alpha}(\nu)$) are critical values of t from the Student's t distribution indexed at the $1-\alpha/2$ (or $1-\alpha$) level and by ν degrees of freedom, and $\mu_1 - \mu_2$ is the difference of the true population means. The computation for hypothetical power is the same as for actual power except that t and ν are recomputed using hypothetical sample sizes instead of the actual sample size.

Two Sample Independent t-Tests

For the two sample independent t-Test, the computed test statistic, $t = \frac{\left(\overline{X}_1 - \overline{X}_2 - d_0\right)}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$, has a Student's t-distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom. For samples i=1 and 2 and data points $j = 1,2,3...n_i$, $\overline{X_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$, d_0 is the estimated difference between sample means, $s^2 = \frac{\left[\left(n_1 - 1\right)s_1^2 + \left(n_2 - 1\right)s_2^2\right]}{\left(n_1 + n_2 - 2\right)}$, and $s_i = \sum_{j=1}^{n_i} \frac{\left(X_{ij} - \overline{X_i}\right)^2}{n_i - 1}$.

For the two sample independent t-Test, the lower and upper limits of a confidence interval for the difference of the sample means are $(\overline{X}_2 - \overline{X}_1) \pm t_{1-\alpha/2}(\nu)\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ where the factor $t_{1-\alpha/2}(\nu)$ is a critical value of t from the Student's t distribution indexed at the $1 - \alpha/2$ level and by ν degrees of freedom. Here alpha (α) is specified by the confidence level where α = 1 - (*confidence level*/100).

Two Sample Paired t-Test

For the two sample paired t-Test, the computed test statistic, $t = \frac{\overline{D} - d_0}{S_D}$, has a Student's t-distribution with $\nu = n - 1$ degrees of freedom (assuming an equal number of n data points in each sample). For samples 1 and 2 and data points $j = 1,2,3...n$, $D = X_{1j} - X_{2j}$, d_0 is the estimated difference between sample means, $\overline{D} = \frac{1}{n}\sum_{j=1}^{n} D_j$, and $S_D = +\sqrt{\frac{1}{n-1}\sum_{j-1}^{n}\left(D_j - \overline{D}\right)^2}$.

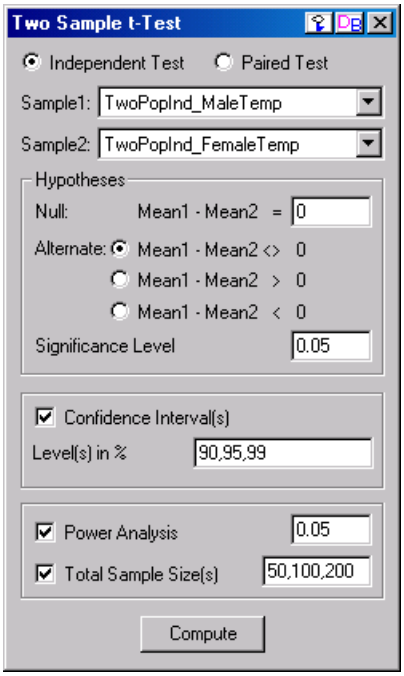
For the two sample paired t-Test, the lower and upper limits of a confidence interval for the difference of the sample means are $(\overline{X}_2 - \overline{X}_1) \pm t_{1-\alpha/2}(\nu)S_D$ where the factor $t_{1-\alpha/2}(\nu)$ is a critical value of t from the Student's t distribution indexed at the $1 - \alpha/2$ level and by ν degrees of freedom. Here alpha (α) is specified by the confidence level where α = 1 - (*confidence level*/100).

Performing a Two Sample t-Test

To perform a two sample t-Test, highlight the desired worksheet columns and then select **Statistics:Hypothesis Testing:Two Sample t-Test**. The menu command opens the Two Sample t-Test dialog box in which you specify the Sample1 and Sample2 data sets as well as all other test parameters

described below. After clicking the Compute button, the test results are output to the Results Log. The results include the type of test performed (Independent or Paired), the names of the sample data sets, sample means ($\overline{X}_1$ and $\overline{X}_2$), standard deviations (SD_1 and SD_2), standard errors (SE_1 and SE_2), and sample sizes (n_1 and n_2). The test statistic (t), degrees of freedom (DF or ν), observed significance (P value), the null (H_0) and alternative (H_1) hypotheses, and the decision rule of the test are output as well. If the Confidence Interval(s) and Power Analysis check boxes are enabled, then all specified confidence intervals and the actual power of the experiment are output. Selecting the Sample Size(s) check box causes a hypothetical power for each specified sample size to also be output.

Operating the Two Sample t-Test Controls



1) The Test Type Radio Buttons

Select the Independent Test or Paired Test radio button depending on the test type you want to conduct. The Paired Test type requires that sample 1 and sample 2 have the same number of data points and that the data points are experimentally paired.

2) The Sample1 and Sample2 Combination Boxes

Select the sample data sets on which to perform a Two Sample t-Test. The sample data sets are assumed to have been drawn from population(s) that follow a normal distribution.

3) The Hypotheses Group

Enter the estimated difference of the sample means, d_0, in the Null Hypothesis edit box to the right of the text "Mean1 - Mean2 =." The value entered for the difference of the sample means is updated in each of the three Alternate Hypotheses.

Select the appropriate radio button for Alternate Hypothesis indicating whether a two-tailed or one-tailed t-Test is to be performed. The direction of the one-tailed t-Test is also specified by the radio button selection. The (in)equality of the Null Hypothesis updates accordingly.

Enter a decimal value greater than 0 and less than 1 for alpha (α) in the Significance Level edit box. The Power Analysis edit box in the Power Analysis group is initialized to the same value of alpha (α) and, if unique, a complementary ([1-α]$*$100) confidence level is added to the Level(s) in % edit box in the Confidence Interval(s) group.

4) The Confidence Interval(s) Group

Select or clear the Confidence Interval(s) check box to determine whether or not confidence intervals are computed. The Level(s) in % edit box is enabled when the Confidence Interval(s) check box is selected.

Enter a comma separated list of confidence levels greater than 0 and less than 100 in the Level(s) in % edit box.

5) The Power Analysis Group

Select or clear the Power Analysis check box to determine whether or not the "actual" power of the t-Test is computed. The Power Analysis edit box (located to the right of its check box) and the Sample Size(s) check box are enabled when the Power Analysis check box is selected.

Enter a decimal value greater than 0 and less than 1 for alpha (α) in the Power Analysis edit box. By default, the Power Analysis edit box is initialized to the same value of alpha (α) that is entered in the Significance Level edit box but it can be modified if desired. The "actual" power computation uses the sample sizes of the selected sample data sets and the alpha (α) specified in the Power Analysis edit box.

Select or clear the Sample Size(s) check box to determine whether or not "hypothetical" powers for the t-Test are computed. The Sample Size(s) edit box (located to the right of its check box) is enabled when the Sample Size(s) check box is selected.

Enter a comma separated list of hypothetical sample sizes (positive Integers only) in the Sample Size(s) edit box. A "hypothetical" power is computed for each hypothetical sample size entered. The sample sizes entered are considered to be total (aggregate) sample sizes for the Independent Test and individual (separate) sample sizes for the Paired Test. The hypothetical power computation uses the alpha (α) specified in the Power Analysis edit box.

6) The Compute Button

Click the Compute button to perform the indicated computations on the selected sample data sets. Control settings (including Sample1 and Sample2) can be changed and the Compute button can be clicked as many times as desired without closing the dialog box. All results are output to the Results Log.

Example: Performing a Two Sample (Independent) t-Test

1) To perform a two sample (independent) t-Test, open the Origin sample project file TWO SAMPLE T-TEST.OPJ located in the Origin \SAMPLES\ANALYSIS\STATISTICS subfolder.

2) Activate the worksheet TwoPopInd and highlight both columns of values. This data file contains 65 observations of male and female normal body temperature (°F). (The column on the left contains the data for men and the column on the right contains the data for women.)

3) Select **Statistics:Hypothesis Testing:Two Sample t-Test**.

4) Select the Independent Test radio button, and clear the Confidence Interval(s), Power Analysis, and Sample Size(s) check boxes.

5) Accept all other default values and click the Compute button.

Origin displays the following results:

[2/11/2002 13:43 "/TwoPopInd" (2452316)]

Two Sample Independent t-Test

Summary Statistics

Sample	N	Mean	SD	SE
1. TwoPopInd_MaleTemp	65	98.10462	0.69876	0.08667
2. TwoPopInd_FemaleTemp	65	98.39385	0.74349	0.09222

Difference of Means: -0.28923

Null Hypothesis: Mean1 - Mean2 = 0
Alternative Hypothesis: Mean1 - Mean2 <> 0

t	DoF	P Value
-2.28543	128	0.02393

At the 0.05 level, the difference of the population means is significantly different than the test difference (0).

With the significance level (or alpha) set at α=0.05, we conclude that male and female normal body temperature are different. The observed significance (P value=0.02393) tells us our probability of obtaining a more extreme t-value by chance alone when the normal body temperatures of males and females are not different. Therefore, if we had set our level of significance at 0.01 for this test, we would have concluded that male and female normal body temperature does not differ.

Note: For a description of how to perform and interpret confidence intervals and power analysis, see "One Sample t-Tests" on page 457.

Example: Performing a Two Sample (Paired) t-Test

1) To perform a two sample (paired) t-Test, open the Origin sample project file TWO SAMPLE T-TEST.OPJ located in the Origin \SAMPLES\ANALYSIS\STATISTICS subfolder.

2) Activate the worksheet TwoPopPair and highlight both columns of values. This worksheet contains observations of two types of percent body fat measurement techniques on 252 men. (One column uses the Brozek body fat technique and the other column ses the Siri body fat technique).

3) Select **Statistics:Hypothesis Testing:Two Sample t-Test**.

4) Select the Paired Test radio button, and clear the Confidence Interval(s), Power Analysis, and Sample Size(s) check boxes.

5) Accept all other default values and click the Compute button.

Origin displays the following results:

[2/11/2002 13:53 "/TwoPopInd" (2452316)]

Two Sample Paired t-Test

Summary Statistics

Sample	N	Mean	SD	SE
1. TwoPopInd_MaleTemp	65	98.10462	0.69876	0.08667
2. TwoPopInd_FemaleTemp	65	98.39385	0.74349	0.09222

Difference of Means: -0.28923

Null Hypothesis: Mean1 - Mean2 = 0
Alternative Hypothesis: Mean1 - Mean2 <> 0

t	DoF	P Value
-12.68577	64	3.97734E-19

At the 0.05 level, the difference of the population means is significantly different than the test difference (0).

With the significance level (or alpha) set at α=0.05, we conclude that the Brozek and Siri percent body fat measurement techniques are different. The observed significance (P value=3.54206E-7) gives the probability of obtaining a test statistic as extreme or more extreme than t by chance alone when the body fat measuring techniques are in fact not different.

Note: For a description of how to perform and interpret confidence intervals and power analysis, see "One Sample t-Tests" on page 457.

One-Way ANOVA

In addition to the one-way analysis of variance (ANOVA), the One-Way ANOVA dialog box includes Levene's test and the Brown-Forsythe test for equal variance. Computations for actual and hypothetical power and for the Bonferroni, Scheffé, and Tukey post hoc means comparisons are also included.

The one-way analysis of variance can be employed to test whether or not two or more populations have the same mean. One-way analysis of variance assumes that the sample data sets have been drawn from populations that follow a normal distribution with constant variance.

The null hypothesis is that the means of all selected data sets are equal. The alternative hypothesis is that the means of one or more selected data sets are different (not equal).

The test statistic for one-way analysis of variance is the ratio $F = \frac{MSBG}{MSE}$, where

$MSBG = \sum_{i=1}^{r} \frac{n_i(\overline{X}_{i.} - \overline{X}_{..})^2}{r-1}$, $MSE = \sum_{i=1}^{r}\sum_{j=1}^{n_i} \frac{\left(X_{ij} - \overline{X}_{i.}\right)^2}{N-r}$, *r* is the number of data sets, n_i is the number of data points in the *i-th* data set, $\overline{X}_{i.}$ is the mean of the *i-th* data set, $\overline{X}_{..}$ is the global mean of all data points (in all data sets), X_{ij} is the *j-th* data point in the *i-th* data set, and N is the total number of all data points.

A P value for the computed F statistic is then obtained from the F-distribution for the degrees of freedom of the numerator $r-1$ and the degrees of freedom of the denominator $N-r$. Large values of F will lead to small P values. If the P value is smaller than a specified significance level alpha, the null hypothesis is rejected and the decision rule states that the means are significantly different. In other words, at least one of the means is significantly different than the others. If the P value is larger than the specified significance level alpha, the null hypothesis is accepted and the decision rule states that the means are not significantly different.

A general reference for Origin's one-way analysis of variance is *Applied Linear Statistical Models*, Neter J., Kutner M., Nachtsheim C., and Wasserman W. (1996), The McGraw-Hill Companies, Inc., Boston, MA. See Sections 16.8 and 16.9.

Tests for equal variance test whether or not two or more population variances (rather than population means) are significantly different. The computation for Levene's test for equal variance is similar to the computation for the one-way analysis of variance. The equations above are precisely followed except that each X_{ij} is replaced with an X'_{ij} where $X'_{ij} = \left(X_{ij} - \overline{X}_{i.}\right)^2$ prior to all calculations. The computation for the Brown-Forsythe test for equal variance is also similar to the computation for the one-way analysis of variance. Again, the equations above are precisely followed except that each X_{ij} is replaced with X''_{ij} where $X''_{ij} = \left|X_{ij} - M_{i.}\right|$ (where $M_{i.}$ is the median of the *i-th* data set) prior to all calculations. See Section 18.2 of *Applied Linear Statistical Models* (1996) for a discussion on tests of equal variance.

Given that an ANOVA experiment has determined that at least one of the population means is statistically different, a post hoc means comparison subsequently compares all possible pairs of population means in the ANOVA experiment to determine which mean (or means) are significantly different. See Sections 17.5, 17.6 and 17.7 of *Applied Linear Statistical Models* (1996) for a detailed discussion on the Bonferroni, Scheffé, and Tukey post hoc means comparisons. In addition, Origin uses the NAG function **nag_anova_confid_interval (g04dbc)** to perfrom means comparisons. See the NAG documentation for more detailed information. (When you install Origin, you are provided the option to install the NAG PDF files which document the NAG functions. If you clicked Yes to install these files, a \NAG PDFs folder is created with a subfolder for each library. If you did not install the PDFs, they remain accessible on your installation CD. Furthermore, you can install them at a later date by running the Origin "Add or Remove Files" program.)

The power of a one-way analysis of variance is a measurement of its sensitivity. Power is the probability that the ANOVA will detect differences in the population means when real differences exist. In terms of

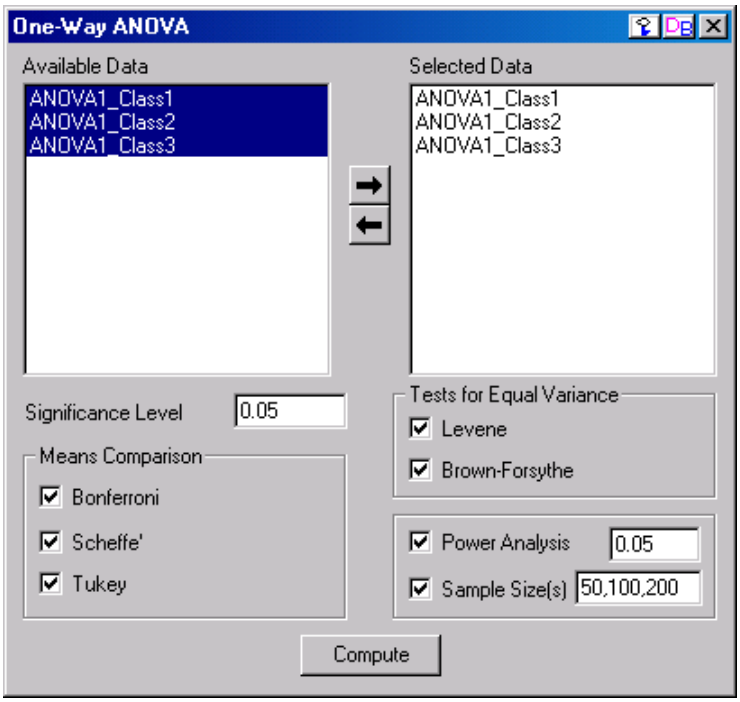
the null and alternative hypotheses, power is the probability that the test statistic F will be extreme enough to allow the rejection of the null hypothesis when it should in fact be rejected (i.e. given the null hypothesis is not true).

Power is defined by the equation:

power = 1 - probf (f , dfa , dfe , nc)

where f is the deviate from the non-central F-distribution with dfa and dfe degrees of freedom and nc = ssa / mse. Ssa is the sum of squares of the Model, mse is the mean square of the Errors, dfa is the degrees of freedom of the numerator, dfe is the degrees of freedom of the Errors. All values (ssa, mse, dfa, and dfe) are obtained from the ANOVA table. The value of probf( ) is obtained using the NAG function **nag_prob_non_central_f_dist**. See the NAG documentation for more detailed information.

Operating the One-way ANOVA Controls



1) The Available Data List Box

Contains a list of all non-Text data sets in the Origin project file that are available to be selected. Highlight the desired available data sets and then click the right-arrow toolbar button to select them.

2) The Selected Data List Box

Contains the selected data sets on which an analysis of variance will be performed. Two or more data sets must be selected before you can perform the ANOVA. De-select selected data sets by highlighting them in the Selected Data list box and then clicking the left-arrow toolbar button.

3) The Significance Level Edit Box

Enter a Significance Level for comparison with the P value when making statistical decisions. The decimal value entered must be greater than 0 and less than 1.

4) The Means Comparison Group

Select or clear the Bonferroni check box determining whether or not the Bonferroni post hoc means comparison is performed.

Select or clear the Scheffé check box determining whether or not the Scheffé post hoc means comparison is performed.

Select or clear the Tukey check box determining whether or not the Tukey post hoc means comparison is performed.

5) The Tests for Equal Variance Group

Select or clear the Levene check box determining whether or not Levene's test for equal variance is performed.

Select or clear the Brown-Forsythe check box determining whether or not the Brown-Forsythe test for equal variance is performed.

6) The Power Analysis Group

Select or clear the Power Analysis check box to determine whether or not the "actual" power of the ANOVA is computed. The Power Analysis edit box (located to the right of its check box) and the Sample Size(s) check box are enabled when the Power Analysis check box is selected.

Enter a decimal value greater than 0 and less than 1 for alpha (α) in the Power Analysis edit box. By default, the Power Analysis edit box is initialized to the same value of alpha (α) that is entered in the Significance Level edit box but it can be modified if desired. The "actual" power computation uses the total sample size of all selected sample data sets and the alpha (α) specified in the Power Analysis edit box.

Select or clear the Sample Size(s) check box to determine whether or not "hypothetical" powers for the ANOVA are computed. The Sample Size(s) edit box (located to the right of its check box) is enabled when the Sample Size(s) check box is checked.

Enter a comma separated list of hypothetical sample sizes (positive Integers only) in the Sample Size(s) edit box. A "hypothetical" power is computed for each hypothetical sample size entered. The sample sizes entered are considered to be total (aggregate) sample sizes. The hypothetical power computation uses the alpha (α) specified in the Power Analysis edit box.

7) The Compute Button

Click the Compute button to perform the indicated computations on the selected sample data sets. Control settings (including selected data sets in the Selected Data list box) can be changed and the Compute button can be clicked as many times as desired without closing the dialog box. All results are output to the Results Log.

Example: Performing a One-way Analysis of Variance

1) To perform a one-way analysis of variance, open the Origin sample project file ONE-WAY ANOVA.OPJ located in the Origin \SAMPLES\ANALYSIS\STATISTICS subfolder.

2) Activate the Anova1 worksheet and highlight all three columns. Select **Statistics:ANOVA:One-Way ANOVA** to open the One-Way ANOVA dialog box. The highlighted data sets are automatically moved into the Selected Data list box.

3) Enter a Significance Level (alpha) of 0.01. Note that the Power Analysis edit box updates and also displays the value 0.01.

4) Select the Bonferroni check box in the Means Comparison group to enable the Bonferroni post hoc means comparison.

5) Select the Levene check box in the Tests of Equal Variance group to enable Levene's test for equal variance.

6) Select the Power Analysis check box to enable the computation of "actual" power for the ANOVA. If desired, enter an alpha of 0.05 for Power Analysis in the associated edit box (or just accept the ANOVA alpha of 0.01).

7) Select the Sample Size(s) check box to enable computation of "hypothetical" powers for the ANOVA. Accept the default hypothetical sample sizes which consists of the comma separated list "50,100,200" (total sample sizes).

8) Click the Compute button to perform the indicated computations.

The One-Way ANOVA is performed. Summary Statistics (including the Null and Alternative hypotheses), an ANOVA table, and a decision rule are output in the Results Log. Similar results for Levene's test for equal variance are also output. Next, the Bonferroni post hoc means comparison is performed and a comparison of each data set's mean against the others is output in the Results Log. Finally, both actual and hypothetical powers for the ANOVA are output.

Two-way ANOVA

Two-way analysis of variance (ANOVA) is a widely used statistical technique for studying the effect that independent (or controlled-by-the-experimenter) variables, called factors, have on a dependent (or response) variable. Two-way analysis of variance experiments have two independent factors each of which have two or more levels. Two-way ANOVA tests for significant differences between the factor level means within a factor and for interactions between the factors. Two-way ANOVA includes several of the basic elements in the Design of Experiments (DOE) which is an important statistical tool in data analysis. In addition to the two-way analysis of variance, the Origin Two-Way ANOVA dialog box

includes computations for the Bonferroni, Scheffé, and Tukey post hoc means comparisons and for actual and hypothetical power.

The Origin Two-Way ANOVA dialog box uses a linear regression approach when computing the two-way analysis of variance and consequently supports unequal sample sizes. This is important because designers of experiments seldom have complete control over the ultimate sample size of their studies. However, Origin's two-way ANOVA does not support empty cells (factor levels with no sample data points). Each of the two factors must have two or more levels and each factor level must have one or more sample data values.

A general reference for Origin's two-way analysis of variance is *Applied Linear Statistical Models*, Neter J., Kutner M., Nachtsheim C., and Wasserman W. (1996), The McGraw-Hill Companies, Inc., Boston, MA. See Section 22.2 of *Applied Linear Statistical Models* for a discussion of how to use dummy variables for testing both factor main effects and interaction effects. In addition, Origin's two-way analysis of variance makes use of several NAG functions. The NAG function **nag_dummy_vars (g04eac)** is used to create the necessary design matrices and the NAG function **nag_regsn_mult_linear (g02dac)** is used to perform the linear regressions of the design matrices. The results of the linear regressions are then used to construct the two-way ANOVA table. See the NAG documentation for more detailed information. (When you install Origin, you are provided the option to install the NAG PDF files which document the NAG functions. If you clicked Yes to install these files, a \NAG PDFs folder is created with a subfolder for each library. If you did not install the PDFs, they remain accessible on your installation CD. Furthermore, you can install them at a later date by running the Origin "Add or Remove Files" program.)

Given that a two-way ANOVA experiment has determined that at least one factor level mean is statistically different than the other factor level means of that factor, a post hoc means comparison subsequently compares all possible pairs of factor level means of that factor to determine which mean (or means) are significantly different. See Sections 17.5, 17.6 and 17.7 of *Applied Linear Statistical Models* (1996) for a detailed discussion on the Bonferroni, Scheffé, and Tukey post hoc means comparisons. In addition, Origin uses the NAG function **nag_anova_confid_interval (g04dbc)** to perform means comparisons. See the NAG documentation for more detailed information.

The power of a two-way analysis of variance is a measurement of its sensitivity. Power is the probability that the ANOVA will detect differences in the factor level means when real differences exist. In terms of the null and alternative hypotheses, power is the probability that the test statistic F will be extreme enough to allow the rejection of the null hypothesis when it should in fact be rejected (i.e. given the null hypothesis is not true).

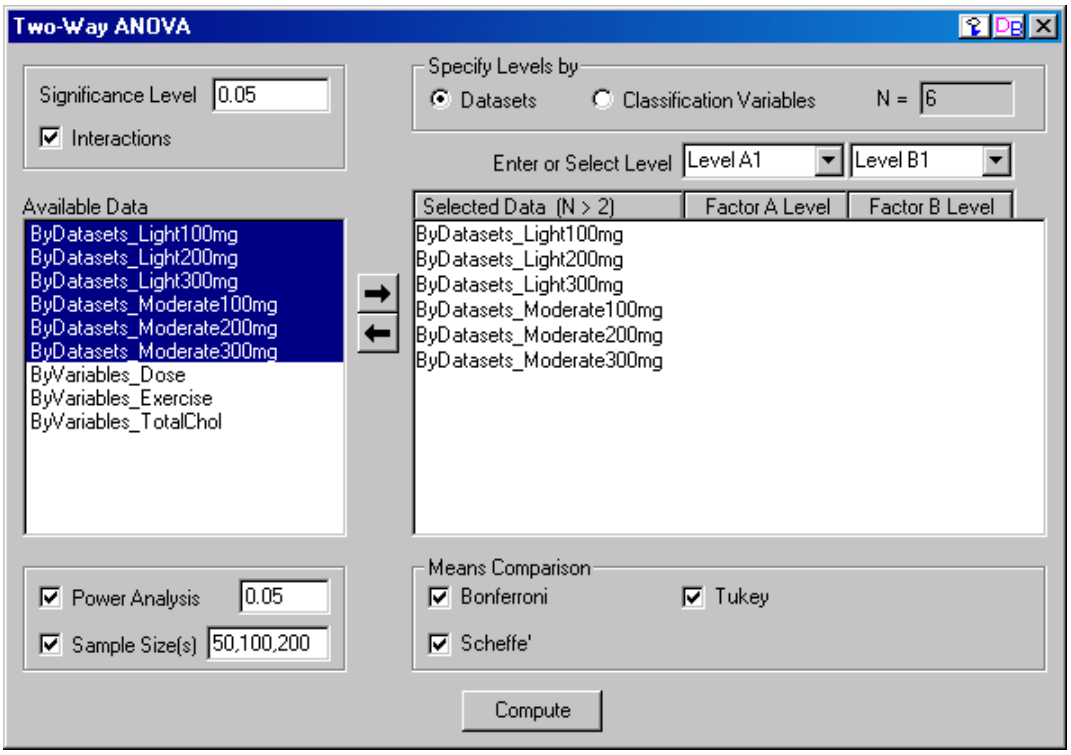
The Origin Two-Way ANOVA dialog box computes power for the Factor A and Factor B sources. If the Interactions check box is selected, Origin also computes power for the Interaction source A * B.

Power is defined by the equation:

power = 1 - probf (f , df , dfe , nc)

where f is the deviate from the non-central F-distribution with df and dfe degrees of freedom and nc = ss / mse. Ss is the sum of squares of the source A, B, or A * B, mse is the mean square of the Errors, df is the degrees of freedom of the numerator for the source A, B, or A * B, dfe is the degrees of freedom of the Errors. All values (ss, mse, df, and dfe) are obtained from the ANOVA table. The value of probf( ) is obtained using the NAG function **nag_prob_non_central_f_dist**. See the NAG documentation for more detailed information.

Operating the Two-way ANOVA Controls



1) The Significance Level Edit Box

Enter a Significance Level for comparison with the P value when making statistical decisions. The decimal value entered must be greater than 0 and less than 1.

2) The Interactions Check Box

Select or clear the Interactions check box determining whether or not interaction effects for the two-way ANOVA are computed.

3) The Specify Levels by Group

Specify how the levels of each factor will be identified. Select the Datasets radio button if the dependent variable data values for each factor level are grouped into separate data sets. Select the Classification Variables radio button if all dependent variable data values are grouped into one data set and their factor levels are identified by two paired classification variable data sets. This setting is predetermined by the manner in which the experimental data is organized.

4) The N = Edit Box

Reports the number of data sets that have been selected. When specifying levels by data sets, two or more data sets must be selected. When specifying levels by classification variables, exactly three data sets must be selected.

5) The Available Data List Box

Contains a list of all data sets in the Origin project file that are available to be selected. Highlight the desired available data sets and then click the right-arrow toolbar button to select them.

6) The Selected Data List Control

Contains the selected data sets on which an analysis of variance will be performed. De-select selected data sets by highlighting them in the Selected Data list control and then clicking the left-arrow toolbar button. The first column of the Selected Data list control always contains the names of all selected data sets. Click the Selected Data column header button to highlight or un-highlight all selected data sets. The remaining columns of the list control are determined by the Specify Levels by radio buttons.

If the Specify Levels by Datasets radio button is selected, then two or more data sets must be selected for the remaining list control columns to properly function. The second column of the list control contains the assigned Factor A Level for each selected data set. Click the Factor A Level column header button to assign the level name displayed in the Factor A Level combo box to all highlighted data sets in the list control. The third column of the list control contains the assigned Factor B Level for each selected data set. Click the Factor B Level column header button to assign the level name displayed in the Factor B Level combo box to all highlighted data sets in the list control. Clicking the Factor A Level or Factor B Level column header button also adds the level name displayed in its combo box to the combo box drop-down list for future selection.

If the Specify Levels by Classification Variables radio button is selected, then exactly three data sets must be selected for the remaining list control columns to properly function. The second column of the list control contains the assigned Variable Type for each selected data set. Click the Variable Type column header button to assign the variable type "Dependent Variable" to the next data set in the list control. There is no third column, but the Switch Factors button < > can be clicked to switch the Factor A and Factor B Classification Variable types.

7) The Factor A Level Comb Box (visible only when specifying levels by data sets)

Enter or select levels in the Factor A Levels drop-down combo box. Entered and selected levels are assigned to the Factor A Level column in the Selected Data list control for all highlighted data sets. Click the Factor A Level column header button to add the displayed level name to the combo box list for future selection.

8) The Factor B Level Combo Box (visible only when specifying levels by data sets)

Enter or select levels in the Factor B Levels drop-down combo box. Entered and selected levels are assigned to the Factor B Level column in the Selected Data list control for all highlighted data sets. Click the Factor B Level column header button to add the displayed level name to the combo box list for future selection.

9) The Power Analysis Group

Select or clear the Power Analysis check box to determine whether or not the "actual" power of the ANOVA is computed. The Power Analysis edit box (located to the right of its check box) and the Sample Size(s) check box are enabled when the Power Analysis check box is selected.

Enter a decimal value greater than 0 and less than 1 for alpha (α) in the Power Analysis edit box. By default, the Power Analysis edit box is initialized to the same value of alpha (α) that is entered in the Significance Level edit box but it can be modified if desired. The "actual" power computation uses the total sample size of all selected sample data sets and the alpha (α) specified in the Power Analysis edit box.

Select or clear the Sample Size(s) check box to determine whether or not "hypothetical" powers for the ANOVA are computed. The Sample Size(s) edit box (located to the right of its check box) is enabled when the Sample Size(s) check box is selected.

Enter a comma separated list of hypothetical sample sizes (positive Integers only) in the Sample Size(s) edit box. A "hypothetical" power is computed for each hypothetical sample size entered. The sample sizes entered are considered to be total (aggregate) sample sizes. The hypothetical power computation uses the alpha (α) specified in the Power Analysis edit box.

10) The Means Comparison Group

Select or clear the Bonferroni check box determining whether or not the Bonferroni post hoc means comparison is performed.

Select or clear the Scheffé check box determining whether or not the Scheffé post hoc means comparison is performed.

Select or clear the Tukey check box determining whether or not the Tukey post hoc means comparison is performed.

11) The Compute Button

Click the Compute button to perform the indicated computations on the selected sample data sets. Control settings (including selected data sets in the Selected Data list control) can be changed and the Compute button can be clicked as many times as desired without closing the dialog box. All results are output to the Results Log.

Example: Performing a Two-Way Analysis of Variance

1) To perform a two-way analysis of variance, open the Origin sample project file TWO-WAY ANOVA.OPJ located in the Origin \SAMPLES\ANALYSIS\STATISTICS subfolder.

2) Activate the ByDatasets worksheet and highlight all six columns. Select **Statistics:ANOVA:Two-Way ANOVA** to open the Two-Way ANOVA dialog box. The highlighted data sets are automatically moved into the Selected Data list control.

3) Enter a Significance Level (alpha) of 0.05. Note that the Power Analysis edit box updates and also displays the value 0.05.

4) Select the Interactions check box to enable the computation of interaction effects.

5) Select the Power Analysis check box to enable the computation of "actual" power for the ANOVA. If desired, enter an alpha of 0.01 for Power Analysis in the associated edit box (or just accept the ANOVA alpha of 0.05).

6) Select the Sample Size(s) check box to enable computation of "hypothetical" powers for the ANOVA. Accept the default hypothetical sample sizes which consists of the comma separated list "50,100,200" (total sample sizes).

7) Select the Tukey check box in the Means Comparison group to enable the Tukey post hoc means comparison.

8) Select the Datasets radio button to specify levels by data set.

9) Drag or CTRL+click to select all the data sets in the Selected Data list control having the word "Light" in their name, so that they become highlighted. Enter "Light" in the Factor A Level combo box and note the Factor A Level setting for the highlighted data sets becomes "Light".

10) Click the Factor A Level column header button and note the item "Light" is added at the bottom of the Factor A Level drop-down combo box.

11) Repeat steps 9 and 10 for the word "Moderate", again using the Factor A Level combo box and column header button. Now repeat steps 9 and 10 for the levels identified by the text "100mg", "200mg", and "300mg", but this time using the Factor B Level combo box and column header button instead of the Factor A Level combo box and column header button.

12) Click the Compute button to perform the indicated computations. The Two-Way ANOVA is performed. Selected data sets and their associated level names and an ANOVA table are output in the Results Log. Next, the Tukey post hoc means comparison is performed on both Factor A and Factor B levels. A comparison of each data set's (level's) mean against the others is output in the Results Log. Finally, both actual and hypothetical powers for the sources of the ANOVA (A, B, and A * B) are output.

13) Click the Selected Data column header button until all selected data sets become highlighted and then click the left-arrow toolbar button removing them from the Selected Data list control.

14) Drag or CTRL+click to select all three data sets in the Available Data list box having the word "ByVariables" in their name, so that they become highlighted. Click the right-arrow toolbar button adding them to the Selected Data list control.

15) Click the Classification Variables radio button to specify levels by classification variables. The Selected Data list control updates and the Factor A and B combo boxes disappear. If the three "ByVariables" data sets are listed in the Selected Data list control, they will automatically be assigned initial variable types. If the three data sets are not selected and the variable types are not automatically assigned, repeat step 14.

16) Click the Variable Type column header button until the data set "ByVariables_TotalChol" is identified as the Dependent Variable. Click the Switch Factors button < > until the data set "ByVariables_Exercise" is identified as the Factor A Classification Variable.

17) Click the Compute button once again to perform the indicated computations. The data in the "ByVariables" worksheet is identical to the data in the "ByDatasets" worksheet, except for its organization. If all other settings remain the same, the new results should be identical to the previously generated results.

Survival Analysis

Survival Analysis is used in bio-science and in quality assurance to quantify survivorship in a population under study. Origin includes two widely used techniques - the Kaplan-Meier Product-Limit Estimator and the Cox Proportional Hazards Model. Both techniques compute a survivorship function (the probability of survival to a given time based on a sample of failure times) and a summary of event (failure) and censored values. In addition, the results of the Kaplan-Meier Product-Limit Estimator include survivorship by quartile with upper and lower confidence limits. The Cox Proportional Hazards Model additionally includes covariate parameter estimates with standard error, chi-square statistic, P value, and hazard ratio.

A general reference for Origin's Survival Analysis features is *Applied Survival Analysis: Regression Modeling of Time to Event Data*, Hosmer, D.W. and Lemeshow, S. (1999), John Wiley and Sons Inc., New York, NY.

Kaplan-Meier Product-Limit Estimator

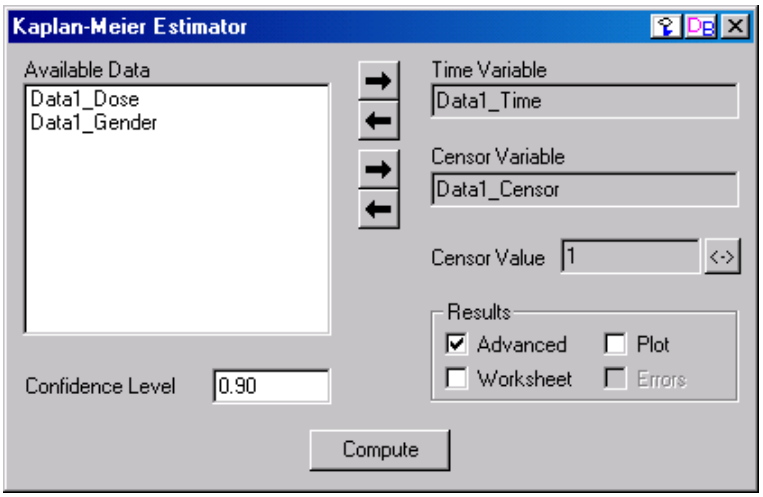
The computational engine for the Origin Kaplan-Meier Product-Limit Estimator dialog box is the NAG function **nag_prod_limit_surviv_fn (g12aac)**. See the NAG documentation for more detailed information. (When you install Origin, you are provided the option to install the NAG PDF files which document the NAG functions. If you clicked Yes to install these files, a \NAG PDFs folder is created with a subfolder for each library. If you did not install the PDFs, they remain accessible on your installation CD. Furthermore, you can install them at a later date by running the Origin "Add or Remove Files" program.)

An additional reference for the Origin Kaplan-Meier Product-Limit Estimator dialog box is Chapter 2 of Applied Survival Analysis: Regression Modeling of Time to Event Data (Hosmer and Lemeshow, 1999).

Cox Proportional Hazards Model

The computational engine for the Origin Cox Proportional Hazards Model dialog box is the NAG function **nag_surviv_cox_model (g12bac)**. See the NAG documentation for more detailed information. An additional reference for the Origin Cox Proportional Hazards Model is Chapter 3 of Applied Survival Analysis: Regression Modeling of Time to Event Data (Hosmer and Lemeshow, 1999).

Operating the Kaplan-Meier Estimator and Cox Proportional Hazards Model



1) The Available Data List Box

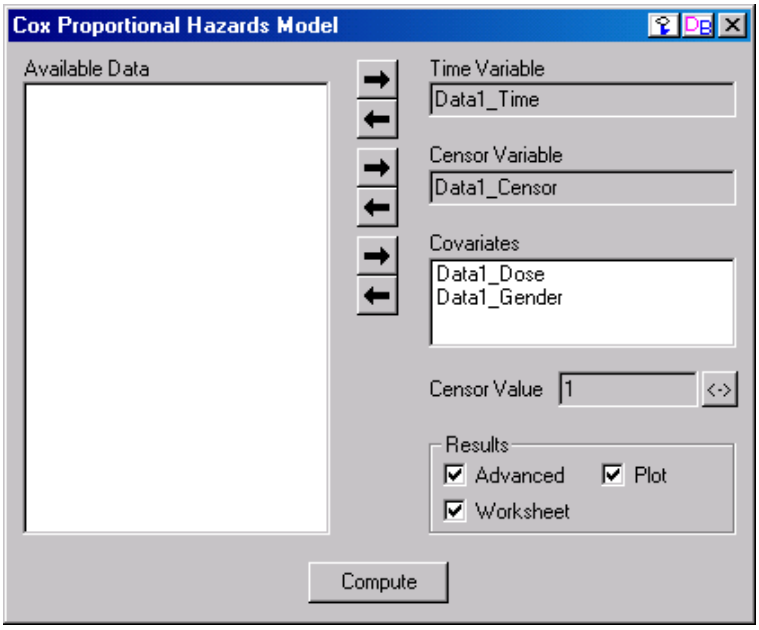
Contains a list of all non-Text data sets in the Origin project file that are available to be selected as Time Variables, Censor Variables, and Covariates (Cox Proportional Hazards Model only). Highlight the desired available data set(s) and then click the appropriate right-arrow toolbar button to select it (them).

2) The Time Variable Edit Box

Contains the selected Time Variable data set. The Time Variable data set is a sample of event (failure) and censored times. Any unit of time may be used but all time values must be positive and must be expressed as elapsed time (relative to the beginning time of the study) and not be absolute time values (calendar dates and clock times). This data set is paired with the Censor Variable data set and must contain precisely the same number of data values. De-select a selected Time Variable data set by selecting a different Time Variable data set or by clicking the appropriate left-arrow toolbar button.

3) The Censor Variable Edit Box

Contains the selected Censor Variable data set. The Censor Variable data set must contain exactly two unique values (text or numeric) one of which identifies a paired time value as an event (or failure) time and one of which identifies a paired time value as censored. The Censor Variable data set must contain precisely the same number of data values as the Time Variable data set. De-select a selected Censor Variable data set by selecting a different Censor Variable data set or by clicking the appropriate left-arrow toolbar button.



4) The Covariates List Box (Cox Proportional Hazards Model only)

Contains one or more selected Covariate data sets. Each Covariate data set is paired with the Censor and Time Variable data sets and must contain precisely the same number of data values. Covariate data set values are Real Numbers but can also be Categorical (e.g. indicate the presence of a condition or characteristic by using the value 1 and the absence of a condition or characteristic by using the value 0; indicate the level of a characteristic by using consecutive integers, etc.). De-select selected Covariate data sets by highlighting them in the Covaiates list box and then clicking the appropriate left-arrow toolbar button.

5) The Censor Value Edit Box

Contains the currently selected Censor Value chosen from one of the two unique values contained in the Censor Variable data set. The Censor Value identifies which time values in the Time Variable data set are censored. An observation having a censored time value represents some outcome other than failure such as prematurely leaving the study due to chance factors or surviving beyond the completion of the study.

6) The Toggle Censor Value Button

Click the Toggle Censor Value Button to choose the Censor Value from one of the two unique values contained in the Censor Variable data set.

7) The Confidence Level Edit Box (Kaplan-Meier Estimator only)

Enter a Confidence Level to be used when computing errors for the survivorship function and for the upper and lower limits of the quartile estimates. The decimal value entered must be greater than 0 and less than 1.

8) The Results Group

Select the Advanced check box if you want the survivorship function output in the Results Log.

Select the Worksheet check box if you want the survivorship function and quartile estimates to be output in a worksheet.

Select the Plot check box if you want a plot of the survivorship function.

Select the Errors check box if you want the errors of the survivorship function to be plotted along with the survivorship function as confidence bands (Kaplan-Meier Estimator only).

9) The Compute Button

Click the Compute button to perform the indicated computations on the selected Time Variable, Censor Variable, and Covariates (Cox only) data sets. All settings can be changed and the Compute button can be clicked as many times as desired without closing the dialog box. Results are output as indicated in the Results Group.

Example: Running the Kaplan-Meier Product-Limit Estimator

1) To run the Kaplan-Meier Product-Limit Estimator, open the Origin sample project file SURVIVAL ANALYSIS.OPJ located in the Origin \SAMPLES\ANALYSIS\STATISTICS subfolder.

2) Activate the Data1 worksheet and select **Statistics:Survival Analysis:Kaplan-Meier Estimator** to open the Kaplan-Meier Estimator dialog box.

3) Highlight the data set **Data1_Time** in the Available Data list box and then click the Select Time Variable data set (right-arrow) toolbar button.

4) Highlight the data set **Data1_Censor** in the Available Data list box and then click the Select Censor Variable data set (right-arrow) toolbar button.

5) Click the Toggle Censor Value button if needed to toggle the selected Censor Value from 0 to 1.

6) Enter a Confidence Level of 0.90.

7) Make sure all four check boxes in the Results Group are selected and then click the Compute button.

The Kaplan-Meier Estimator runs, outputting the Summary of Event and Censor Values, the Survivorship Function, and Quartile Estimates in the Results Log and in a Survival worksheet. The Survivorship Function and Confidence Bands (upper and lower confidence limits) are plotted in a SurvivalPlot graph.

Example: Running the Cox Proportional Hazards Model

1) To run the Cox Proportional Hazards Model, open the Origin sample project file SURVIVAL ANALYSIS.OPJ located in the Origin \SAMPLES\ANALYSIS\STATISTICS subfolder.

2) Activate the Data1 worksheet and select **Statistics:Survival Analysis:Cox Proportional Hazards Model** to open the Cox Proportional Hazards Model dialog box.

3) Highlight the data set **Data1_Time** in the Available Data list box and then click the Select Time Variable data set (right-arrow) toolbar button.

4) Highlight the data set **Data1_Censor** in the Available Data list box and then click the Select Censor Variable data set (right-arrow) toolbar button.

5) Highlight the data sets **Data1_Gender** and **Data1_Dose** in the Available Data list box and then click the Select Covariate data sets (right-arrow) toolbar button.

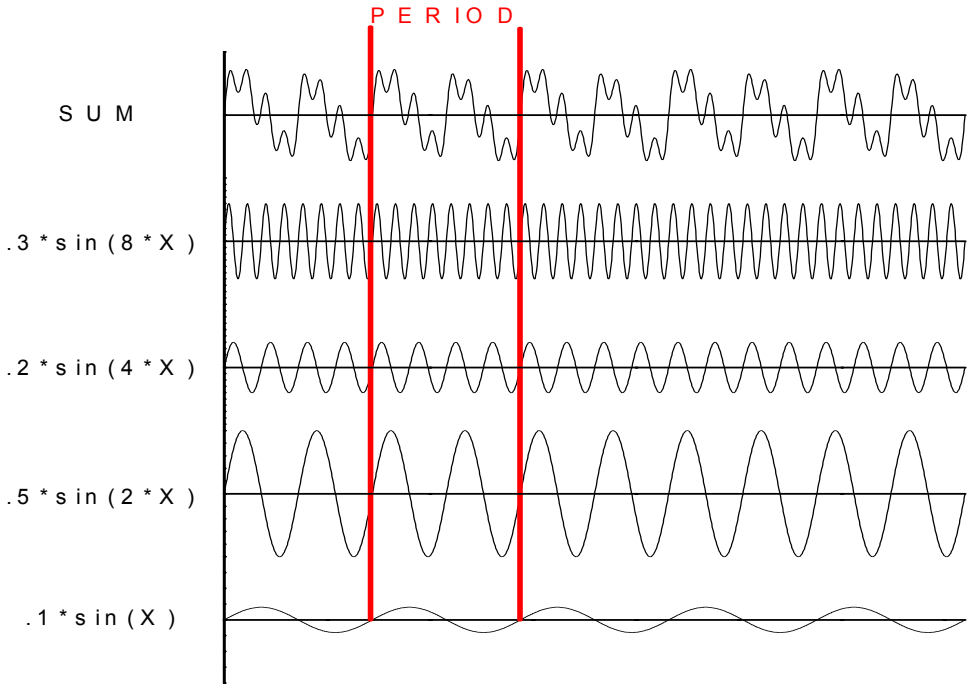
6) Click the Toggle Censor Value button if needed to toggle the selected Censor Value from 0 to 1.

7) Make sure all three check boxes in the Results Group are selected and then click the Compute button.

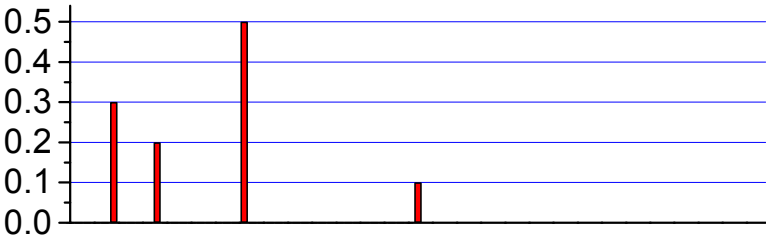
The Cox Proportional Hazards Model runs, outputting the Summary of Event and Censor Values, the Survivorship Function, and Parameter Estimates in the Results Log and in a Survival worksheet. The Survivorship Function is plotted in a SurvivalPlot graph.

Fast Fourier Transform (FFT)

The mathematician Fourier recognized that a periodic function could be described as an infinite sum of periodic functions. In particular, he described the formulas for transforming such periodic functions into sums of harmonics of Sine or Cosine functions. Here is an example of four Sine waves that sum to make a periodic function:



The Fourier Transform would look something like this:



Which shows the amplitude values of specific SINE 'frequencies'.

Collectively, the various transform equations are know as Fourier Transform equations. While common use has been to apply these transforms to functions of Time, there is nothing in the equations that implies such a limitation. The transform is referred to as an inverting function in that the units are inverted. Thus,

data as a function of Time would be transformed to data as a function of 1/Time (frequency). Similarly, data as a function of oranges/acre would be transformed to data as a function of acres/orange. Often the units are referred to as domains, and the transform is said to display time domain data in the frequency domain (for example).

Real world use of these transforms considers discrete data points rather than continuous functions. The data is sampled at regular periods (the sampling rate or interval) over the interval at which the data repeats (the sampling period). The equations or algorithms for these calculations are called Discrete Fourier Transforms (DFT). The inverse of the sampling period represents a lower limit to the resolution of the transformed data. The inverse of the sampling interval would thus seem to represent an upper limit to the resolution of the transformed data, but - due to the Nyquist sampling theorem - the upper limit is half that. For example:

Measurements are taken at 0.001 second intervals for a period of 0.5 seconds. Since seconds are the period units, then Hz (1/sec.) are the transform units. The lower limit of resolution is 1/0.5 or 2 Hz. The upper limit of resolution is 1/2 of 1/0.001 or 500 Hz. We should therefore be able to examine the frequencies in the data in steps of 2 Hz up to 500 Hz.

The Real component of a Fourier transform is symmetrical about the center while the Imaginary component exhibits negative symmetry about the center. The Fourier zero value is variously referred to as the 'Zero Frequency' value, the DC component, and the Mean or Average value.

The time it takes to calculate the DFT increases exponentially as more data points are considered. For the past 50 years or so, mathematicians have exploited redundancies and symmetries in the DFT to reduce computation time. The results of their efforts are collectively called Fast Fourier Transforms (FFT). The fastest of these FFTs are based on equations when the number of data points happens to be an integral power of 2.

Origin and the FFT

The first step in Origin's calculation of the FFT is to make the number of data points an integral power of 2. Origin does this by extending the X data set using the existing step values of X until there are 2^N points and setting all the new Y values to zero, or by truncating the length of a data set until there are 2^N points. The changed sample period results in a slight change in resolution, while the same sampling interval yields the same maximum. The algorithm is based on the initial data set size as follows:

Data size	Excess Power of Two	Action
0 to 511	Any	Always pad to next power of two
512 to 2047	less than 5% over	Truncate - otherwise pad
2048 to 8192	less than 10% over	Truncate - otherwise pad
8192 to 16383	less than 20% over	Truncate - otherwise pad
>16383	less than 30% over	Truncate - otherwise pad

For example:

Given	Next lowest 2^N	Percent cutoff point	Action
2252 points	2048	2048+.1*2048=2252.8	Truncate to 2048
2253 points	2048	2048+.1*2048=2252.8	Pad to 4096

If you want to use interpolation to first get an exact power of two, then you must first plot the data as a line plot and then use the **Analysis:Interpolate/Extrapolate** menu command to generate an equal-spaced data set of however many 2^N points you need. Use the FFT on the new data set.

The FFT Tool presents a number of options for performing an FFT, including Forward, Backward, Amplitude, Power, Normalization, Shifting and Phase Unwrapping as well as four 'windowing' functions for enhancing resolution. Some of these options only affect the display of data and some affect magnitudes - in particular, use of Windowing options include scaling factors which are not easily removable or reversible. While the Fourier Transform is its own inverse, most practical implementations of the DFT and the FFT include a sign difference in the imaginary part of its results. Origin makes this difference a selectable option.

The Origin FFT calculates the Real, Imaginary, Complex , Phase and Power components. The plots used by the Forward FFT are (by convention) vs. Frequency. The plots used by the Backward FFT are vs. Time. The complex result in the 'r' column is derived from the square root of the sum of the squares of the Real and Imag(inary) columns. The phase angle in the 'Phi' column is derived from the angle whose tangent is Imag/Real with corrections for quadrant based on the phase unwrap option. The Power is derived from the Real^2/N, where N is the number of points.

There is no difference in the worksheets created by selecting either Amplitude or Power. The complex or 'r' column is displayed in plots of Amplitude while the Power column is displayed in plots of Power.

Normalization in a Forward FFT divides the Real results by N/2, except for the DC component which is divided by N. Normalization in a Backward FFT divides the Real results by 2, except for the DC component which is left unchanged.

Shifting will change the X scale and adds an additional data point that duplicates the first (completing a natural symmetry). The unshifted results are 0 to a maximum, while the shifted results are between plus and minus a half-maximum.

By Phase wrapping, Origin keeps the phase angle between plus and minus 180 degrees. Unwrapping the phase allows the angle to vary outside those limits based on the sign changes of the real and imaginary columns.

For more information, go to www.originlab.com and search the Knowledge Base for “FFT”.

Practical Problems With the FFT

Problem: After selecting **Analysis:FFT**, an Attention box opens displaying the message “Sampling resolution test failed! Please check your data and set the sampling interval in FFT settings.”

Response: This error message displays when the sampling rate (interval between each X or time value) changes. Among its other restrictions, the FFT requires that sampling data be equally spaced. Although the FFT algorithm will operate on data not equally spaced, it will produce incorrect results.

A small change in sampling rate could be due to instrument instability during data collection. This kind of fluctuation will not have a significant effect on the results of the FFT. However a larger change in sampling rate, such as a data spike, will produce incorrect results.

The sampling interval used to perform the FFT is calculated using only the first few cells of X data. If there is a spike in this data range, it can result in the sampling rate being set incorrectly on the FFT tool's Settings tab. If this is the case, type the correct sampling rate value in the Sampling Interval text box on the Settings tab before performing the FFT. This will allow for accurate results to be calculated.

Problem: When I do an FFT of a known SINE wave, I don't get a perfect amplitude peak at my known frequency.

Response: When you look at the result of an FFT, you can think of it as looking at the 'true' FFT through a picket fence. Changing the picket spacing or shifting your viewpoint will give you a slightly different, but equally valid view. The number of points in the original data set and the sample period will have analogous effects on the FFT. To get the ideal situation asked for, the sample would have to represent one (or more) complete periods with exactly 2^n points of data.

Problem: When I do a Backward FFT of my Forward FFT, I don't get my original data.

Response: The settings for Origin 3.x were such that clicking the Inverse FFT button on the FFT Template returned the original data set (plus any padded points). The correct procedure in 4.x through 7.x is as follows:

1) Do the Forward FFT with Normalization and Shift OFF (unchecked).

The Complex result will be plotted along with Phase. You can modify the Real and Imag values in the FFT worksheet to do Frequency Domain Filtering.

2) Select Real and Imag from the FFT results worksheet and do the Backward FFT with Normalization and Shift OFF (unchecked).

The Complex result will be plotted along with Phase.

3) The Real column of the IFFT results worksheet will be the same as the original data set (plus any padded zero values).

NAG FFT

Origin 7 includes a number of the NAG function libraries, including c06 - Fourier Transforms. Reference information is provided on the NAG libraries in the Origin C Reference Help file (**Help:Programming:Origin C Reference**). Furthermore, when you install Origin, you are provided the option to install the NAG PDF files which document the NAG functions. If you clicked Yes to install these files, a \NAG PDFs folder is created with a subfolder for each library. If you did not install the

PDFs, they remain accessible on your installation CD. Furthermore, you can install them at a later date by running the Origin "Add or Remove Files" program.

To help you get started calling NAG functions from Origin C functions, a number of sample Origin projects and associated source files are included in your Origin \SAMPLES\PROGRAMMING subfolders. Those related to FFT include:

\NAG 2D FFT\

\NAG FFT Convolution\

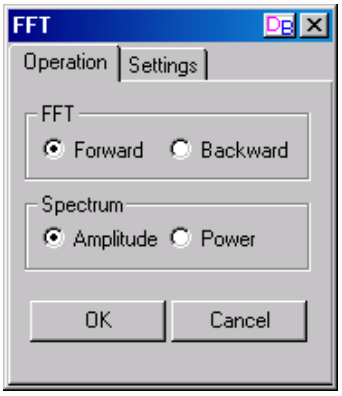
\NAG FFT Lowpass\

\NAG STFT\

Using the FFT Tool

To perform a fast Fourier transform (FFT) on highlighted data in the worksheet or on the active data plot, select **Analysis:FFT**. This menu command opens the FFT tool that allows you to prepare the data and choose the real, imaginary, and time components for the calculation. Origin's FFT assumes that your independent variable (X data set) is time and that your dependent variable (Y data set) is some sort of amplitude.

The Operation Tab



The FFT Group

Select the Forward radio button to perform a forward FFT calculation when you click OK.

Select the Backward radio button to perform a backward FFT calculation when you click OK.

The Spectrum Group

Select the Amplitude radio button to plot amplitude and phase data.

Select the Power radio button to plot the power spectrum and phase data.

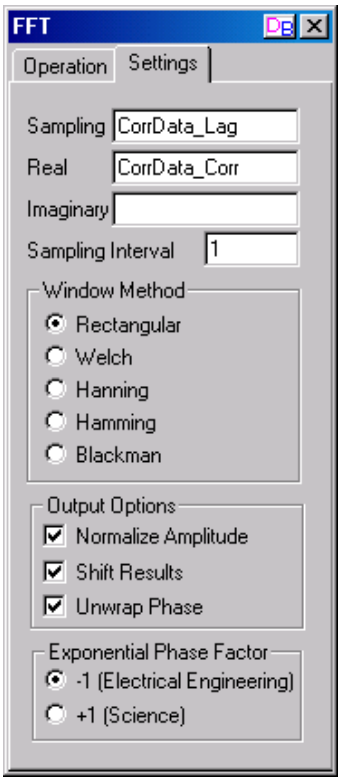
The OK Button

Click OK to perform an FFT. The resulting data is displayed in a new worksheet and graph window.

The Cancel Button

Click Cancel to close the tool.

The Settings Tab



Select this tab to access the Settings options.

The Sampling Text Box

Type a data set to be used for time or frequency information. By default, the X data set of the active data plot (or the worksheet's X column) is selected as the sampling data set.

The Real Text Box

Type a data set to be used as the real component for the FFT calculation. By default, the Y data set of the active data plot is selected as the real data set.

The Imaginary Text Box

Type a data set to be used as the imaginary component for the complex FFT calculation. If no data set is specified, a real FFT will be performed.

The Sampling Interval Text Box

Type the time or frequency interval to be used in the FFT calculation. Increase this value if a Time Resolution error occurs.

The Window Method Group

Select the window method to be used in the FFT calculation.

1) Rectangular Window: $w[n] = 1$ for $0 \le n \le N - 1$ and zero otherwise. This is the window function used in pre-4.0 versions of Origin.

Using appropriate window functions other than the rectangular window can enhance the spectrum resolution.

2) Welch Window: $$w[n] = 1 - \left(\frac{n - \frac{1}{2}(N-1)}{\frac{1}{2}(N+1)} \right)^2$$

3) Hanning Window: $$w[n] = \frac{1}{2}\left[1 - \cos\left(\frac{2\pi n}{N-1} \right) \right]$$

4) Hamming Window: $$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1} \right)$$

5) Blackman Window: $$w[n] = 0.42 - 0.5\cos\left(\frac{2\pi n}{N-1} \right) + 0.08\cos\left(\frac{4\pi n}{N-1} \right)$$

The Output Options Group

1) Select the Normalize Amplitude check box to perform amplitude normalization. The effect on the FFT result is to divide the amplitudes of the DC and AC components by N/2, where N is the number of data points. This will reveal the true amplitudes in the original data set. This occurs because we know that $\cos(x) = [\exp(-ix) + \exp(ix)]/2$. Thus, when a time domain data set is transformed into the frequency domain by FFT, each component splits into two frequencies, a positive one and its negative image. The amplitude of each of these frequencies is N/2 times that of its original component.

To calculate the mean of the data set, divide the DC component by 2.

2) The Shift Results check box determines how the transformed data will be presented. If the check box is cleared, the transform is displayed for positive frequencies only, similar to displaying the phase in the range of 0° to 360°. This is in better agreement with the normal definition of DFT. If the check box is

selected, the transform is shifted around to be displayed with both positive and negative frequencies centered at zero, similar to displaying the phase in the range of –180° to +180°. There is one extra frequency point involved in this presentation. Some symmetrical properties of FFT can be better seen in this form.

3) Select the UnWrap Phase check box to keep the phase unwrapped, to maintain the true phase data. By default, phase data is wrapped so that it is defined in the range of –180° to +180°.

The Exponential Phase Factor Group

Select the Electrical Engineering or Science convention to set the sign of the Exponential Phase factor for the FFT operation. If you select the Science option, the phase factor will be set according to the formulae listed in page 503 of Numerical Recipes in C, 2nd edition. If you select the Electrical Engineering option, the phase factor will be of opposite sign compared with the Science option. The two definitions give the same real components, but their imaginary components and the phase angle will have opposite signs.

The FFT Tool Results

After clicking OK on the Operation tab of the FFT tool, Origin performs the FFT and appends the results to two new windows: a graph and worksheet window.

The FFTPlot*n* graph window displays the amplitude and phase information for the transformed data. Because Origin's FFT assumes that the independent variable (X data set) is time (in seconds) and that the dependent variable (Y data set) is some sort of amplitude, the X axis for the FFT graph window will be scaled in units of Hz. If your data is not time-series data, you will need to re-label your axis.

The FFT worksheet window contains the frequency data, the real and imaginary parts of the transformed data, the polar form of the transformed data, and the power spectrum data.

The frequency column is obtained from the time data set as follows: If the time separation between successive abscissas is Δt, then the n^{th} frequency datum is:

$$f_n = \frac{n}{N\Delta t}$$

If the time data are given by the row index, then Δt is simply unity.

If there are N input data points, the frequency domain will also have N points with the maximum frequency, $f_{\max}$ equal to $\frac{1}{\Delta t}\left(1-\frac{1}{N}\right)$ where Δt is the time step between points. If the Shift Results check box on the Settings tab of the FFT tool has been cleared, the unshifted transform displays from 0 to $f_{\max}$. Otherwise, the shifted transform displays from $-\frac{f_{\max}}{2}$ to $\frac{f_{\max}}{2}$.

Performing a Backward FFT

Origin allows you to apply a backward (or inverse) FFT on the FFT result to transform back the original data set. Thus, you can apply a backward FFT on the Freq(X), Real(Y), and Imag(Y) data sets in the FFT results worksheet. In principle, this should transform the FFT result back to its original data set. However, this claim is valid only for the FFT results when both the Normalize Amplitude and Shift Results check boxes are cleared when first applying FFT to the (original) data set. Otherwise, if the Normalize Amplitude check box is selected when performing the FFT, the amplitude ratio between AC and DC components will be distorted by a factor of 2. Additionally, if the Shift Results check box is selected when performing the FFT, the FFT results will be in a computationally wrong order and, furthermore, the total number of frequency points is incorrect. Consequently, in this situation the backward FFT will not get the FFT result back to its original data.

FFT Mathematical Descriptions

Discrete Fourier Transform (DFT)

For a data set $x[n]$ with index n in the range $0 \le n \le N-1$, ($x[n]$ can be real or complex numbers), the (forward) Discrete Fourier Transform (DFT) is defined to be:

$$X[k] = \sum_{n=0}^{N-1} x[n] \exp(-i2\pi F_k n) \qquad 0 \le n \le N-1$$

where $F_k = \frac{k}{N}$. Note that DFT transforms N complex numbers $x[n]$ (real numbers are complex numbers with imaginary components having zero values) into N complex numbers $X[k]$. Also note that DFT involves only the data values and their indices. Other variables associated with the data (such as time) are not needed in the calculation. In practice, DFT is often performed on data collected at an equal (time) interval τ. It is easy to convert the index into "time" $t = n\tau$, and $F_k = \frac{k}{N}$ into "frequency" $f_k = \frac{k}{N\tau}$.

The inverse (backward) Discrete Fourier Transform (DFT) is defined to be:

$$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k] \exp(i2\pi F_k n) \qquad 0 \le k \le N-1$$

The formulae for forward and inverse DFT presented here follow the phase convention used in Electrical Engineering. It differs from the definition in Numerical Recipes in the sign of the phase factor in the exponent. The FFT definition in Numerical Recipes follows:

$$X[k] = \sum_{n=0}^{N-1} x[n] \exp(i2\pi F_k n) \qquad 0 \le n \le N-1$$

$$x[n]=\frac{1}{N}\sum_{k=0}^{N-1}X[k]\exp(-i2\pi F_k n)\quad 0\le k\le N-1$$

The two definitions give the same real components, but their imaginary components have the opposite sign. You can select between the Engineering and Science phase conventions in Origin under the Settings tab of the FFT interface.

Parseval's theorem can be used to check the results of DFT:

$$\sum_{n=0}^{N-1}\left|x[n]\right|^2=\frac{1}{N}\sum_{k=0}^{N-1}\left|X[k]^2\right|$$

Fast Fourier Transform (FFT)

Every efficient algorithm of DFT can be called an FFT. Origin implements the Danielson-Lanczos method (see references). Danielson and Lanczos showed that if the total number of data points N is an integer power of 2, the DFT of these N numbers can be rewritten as the sum of two DFTs, each of length $N/2$:

$$X[k]=\sum_{j=0}^{N/2-1}x[2j]\exp\left(-i\frac{2\pi kj}{(N/2)}\right)+W^k\sum_{j=0}^{N/2-1}x[2j+1]\exp\left(-i\frac{2\pi kj}{(N/2)}\right)$$

where $W=\exp\left(-i\frac{2\pi}{N}\right)$. We can use the Danielson-Lanczos Lemma recursively, thus reducing the calculation to ~ $N\log_2 N$, compared to ~ N^2 required by DFT.

Power Spectrum Estimation Using FFT

The FFT of a set of sampled data is not the true Fourier transform of the process from which the data was obtained. The discrete data is obtained effectively by a window function $w[n]$ which selects a finite-length (N samples) of $x[n]$ from the continuous signal:

sampled data $v[n]=w[n]x[n]$

($w[n]=0$ for *n* outside the range $0\le n\le N-1$)

The FFT of the sampled data is:

$$V[k]=\sum_{n=0}^{N-1}w[n]x[n]\exp(-i2\pi F_k n)\qquad 0\le n\le N-1$$

An estimate of the power spectrum is:

$P[k]=\frac{1}{\sum_{n=0}^{N-1}(w[n])^2}\left|V[k]\right|^2$, which is also called periodogram.

Some of the popular window functions are listed below:

1) Rectangular Window: $w[n] = 1$ for $0 \le n \le N-1$ and zero otherwise. This is the window function used in pre-4.0 versions of Origin.

Using appropriate window functions other than the rectangular window can enhance the spectrum resolution.

2) Welch Window: $$w[n] = 1 - \left(\frac{n - \frac{1}{2}(N-1)}{\frac{1}{2}(N+1)} \right)^2$$

3) Hanning Window: $$w[n] = \frac{1}{2}\left[1 - \cos\left(\frac{2\pi n}{N-1} \right) \right]$$

4) Hamming Window: $$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1} \right)$$

5) Blackman Window: $$w[n] = 0.42 - 0.5\cos\left(\frac{2\pi n}{N-1} \right) + 0.08\cos\left(\frac{4\pi n}{N-1} \right)$$

Correlation using the FFT

To measure the correlation between two columns of data, highlight the desired data (independent of column designation) and select **Analysis:Correlate**. Origin adds two columns to the rightmost position in the worksheet. The left column holds the resultant lag or index variables, and the right column holds the correlation result.

The absolute value of the correlation result will be large when the leftmost data set is exactly shifted to the right or to the left of the second data set by the lag value.

To learn more about correlation, review the FFT CORRELATION.OPJ project located in your Origin \SAMPLES\ANALYSIS\FFT folder.

Convolution using the FFT

The convolution of two data sets is a general process that can be used for various types of data smoothing, signal processing, or edge detection. The leftmost data set, the signal data set, is convolved by the second data set, the response data set. The response data set should meet the following requirements:

1) The response data set should consist of an odd number of points and be a representative sample of a symmetric function.

2) The number of points, r, in the response data set must be less than half the number of points, s, in the signal data set. (The last r points, and, to a lesser extent, the first r points, of the s points in the result are of no value. Therefore, 2*r should be much less than s.)

3) The sum of the points in the response curve should be unity in order to retain the amplitude of the original data set. Otherwise the convolution result will be scaled by a factor equal to the sum.

To avoid possible artifacts from the FFT (performed as part of the convolution process), you should pad the signal data with zero values until you have an integral power of two number of points. You will also need to extend your X data accordingly.

To convolve a data set with a response data set, highlight both data sets and select **Analysis:Convolute**. Origin adds two columns to the rightmost position in the worksheet. The left column holds the index variables, and the right column holds the convolution result.

To learn more about convolution, review the FFT CONVOLUTION.OPJ project located in your Origin \SAMPLES\ANALYSIS\FFT folder.

Note: Origin 7 includes a number of the NAG function libraries, including c06 - Fourier Transforms. Reference information is provided on the NAG libraries in the Origin C Reference Help file (Help:Programming:Origin C Reference). Furthermore, when you install Origin, you are provided the option to install the NAG PDF files which document the NAG functions. If you clicked Yes to install these files, a \NAG PDFs folder is created with a subfolder for each library. If you did not install the PDFs, they remain accessible on your installation CD. Furthermore, you can install them at a later date by running the Origin "Add or Remove Files" program.

To help you get started calling NAG functions from Origin C functions, a number of sample Origin projects and associated source files are included in your Origin \SAMPLES\PROGRAMMING subfolders. Those related to FFT include:

\NAG 2D FFT\

\NAG FFT Convolution\

\NAG FFT Lowpass\

\NAG STFT\

Deconvolution using the FFT

Deconvolution is a process that can undo the “blurring” obtained after convoluting data. While convolution is the product of the signal and response data sets, deconvolution is achieved by dividing the known convolution by the response data set.

The response data set should meet the following requirements:

1) The response data set should consist of an odd number of points and be a representative sample of a symmetric function.

2) The number of points, r, in the response data set must be less than half the number of points, s, in the signal data set. (The last r points, and, to a lesser extent, the first r points, of the s points in the result are of no value. Therefore, 2*r should be much less than s.)

3) The sum of the points in the response curve should be unity in order to retain the amplitude of the original data set.

To avoid possible artifacts from the FFT (performed as part of the deconvolution process), you should pad the signal data with zero values until you have an integral power of two number of points. You will also need to extend your X data accordingly.

To deconvolve a data set with a response data set, highlight both data sets (the convolved data set should be on the left, the response data set on the right) and select **Analysis:Deconvolute**. Origin adds two columns to the rightmost position in the worksheet. The left column holds the index variables, and the right column holds the deconvolution result.

To learn more about deconvolution, review the FFT DECONVOLUTION.OPJ project located in your Origin \SAMPLS\ANALYSIS\FFT folder.

Data Smoothing and Filtering

Origin provides the following data smoothing and filtering options:

1) Smoothing using Savitzky-Golay filtering.
2) Smoothing using adjacent averaging.
3) FFT filter smoothing.
4) Digital filtering using low pass, high pass, band pass, band block, and threshold filters.

The three smoothing methods are available from the **Analysis:Smoothing** menu, or from the Smoothing tool (**Tools:Smooth**). The smoothed data is placed in a newly created (hidden) worksheet named Smoothed*n*. The window label reports the type of smoothing that was performed. Additionally, the smoothed data plot is added to the active layer of the graph. The Smoothing tool allows you to replace the original data instead of creating a new worksheet.

Smoothing using Savitzky-Golay Filtering

To smooth the active data plot using the Savitzky-Golay filter method, select **Analysis:Smoothing:Savitzky-Golay**. This menu command opens the Smoothing dialog box. The degree of the underlying polynomial has a default value of 2, with an upper limit of 9. This parameter allows you to improve the fitting/differentiation. To alter this value, select the desired value from the drop-down list. (For an illustration of the effect of this parameter on smoothing, see *Numerical Recipes in C* by Press et. al., Second Edition, Fig. 14.8.2, page 654.) Origin includes a function to calculate a large number of Savitzky Golay coefficients, allowing you to have large windows (~ 100 points). Further, you can make the windows asymmetric about a particular data point. For example, the number of points to the left of the point of interest can be different from the number to the right. Make your selection from the associated drop-down lists.

The Savitzky-Golay filter method essentially performs a local polynomial regression to determine the smoothed value for each data point. This method is superior to adjacent averaging because it tends to preserve features of the data such as peak height and width, which are usually 'washed out' by adjacent averaging.

Smoothing using Adjacent Averaging

To smooth the active data plot by averaging adjacent data points, select **Analysis:Smoothing:Adjacent Averaging**. This menu command opens the Smoothing dialog box. Specify a number that controls the degree of smoothing. If you enter an odd number *n,* then *n* points are used to calculate each averaged result. If you enter an even number *m,* then *m*+1 points are used to calculate each averaged result. The smoothed value at index *i* is the average of the data points in the interval [*i*-(*n*-1)/2, *i*+(*n*-1)/2], inclusive.

FFT Filter Smoothing

To smooth the active data plot by FFT filtering, select **Analysis:Smoothing:FFT Filter**. This menu command opens the Smoothing dialog box in which you specify how many data points at a time to be considered by the smoothing routine.
The smoothing is accomplished by removing Fourier components with frequencies higher than:

$$\frac{1}{n\Delta t}$$

where *n* is the number of data points considered at a time, and Δt is the time (or more generally the abscissa) spacing between two adjacent data points.

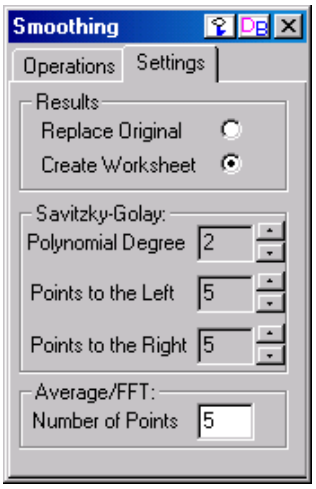
Note 1: The function used to clip out the high-frequency components is a parabola with its maximum of 1 at zero frequency and falling to zero at the cutoff frequency defined above. The parameters of this parabolic clipping function are determined by the total number of points and the number of points considered at one time. The more points considered at one time, the greater the degree of smoothing. A value of zero for this parameter will leave the data unsmoothed.

Note 2: The various FFT filtering techniques (Low/High Pass, Band Pass, etc.) are not affected by data masking.

Using the Smoothing Tool

To open the Smoothing tool, make the desired data plot active and select **Tools:Smooth**.

The Settings Tab



Click this tab to access the Settings options.

The Results Group

Select the Replace Original radio button to replace the raw data plot with the smoothed data.

Select the Create Worksheet radio button to create a new hidden worksheet containing the smoothed data.

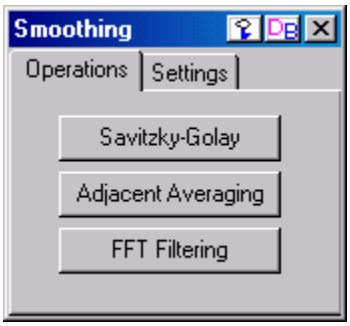
Additional Controls

The degree of the underlying polynomial (Polynomial Degree) has a default value of 2, with an upper limit of 9. This parameter allows you to improve the fitting/differentiation. Select the desired value from the associated control. (For an illustration of the effect of this parameter on smoothing, see *Numerical Recipes in C* by Press et. al., Second Edition, Fig. 14.8.2, page 654.)

Origin includes a function to calculate a large number of Savitzky Golay coefficients, allowing you to have large windows (~ 100 points). Further, you can make the windows asymmetric about a particular data point. For example, the number of points to the left of the point of interest can be different from the number to the right. Make your selection from the Points to the Left (and Right) controls.

For adjacent averaging and FFT filter smoothing, specify the number of data points at a time to be considered by the smoothing routine in the associated text box.

The Operations Tab



Click on this tab to access the Operations options.

The Savitzky-Golay Button

Click this button to perform Savitzky-Golay filter smoothing calculations on the active data plot. The settings on the Settings tab control the degree of smoothing.

The Adjacent Averaging Button

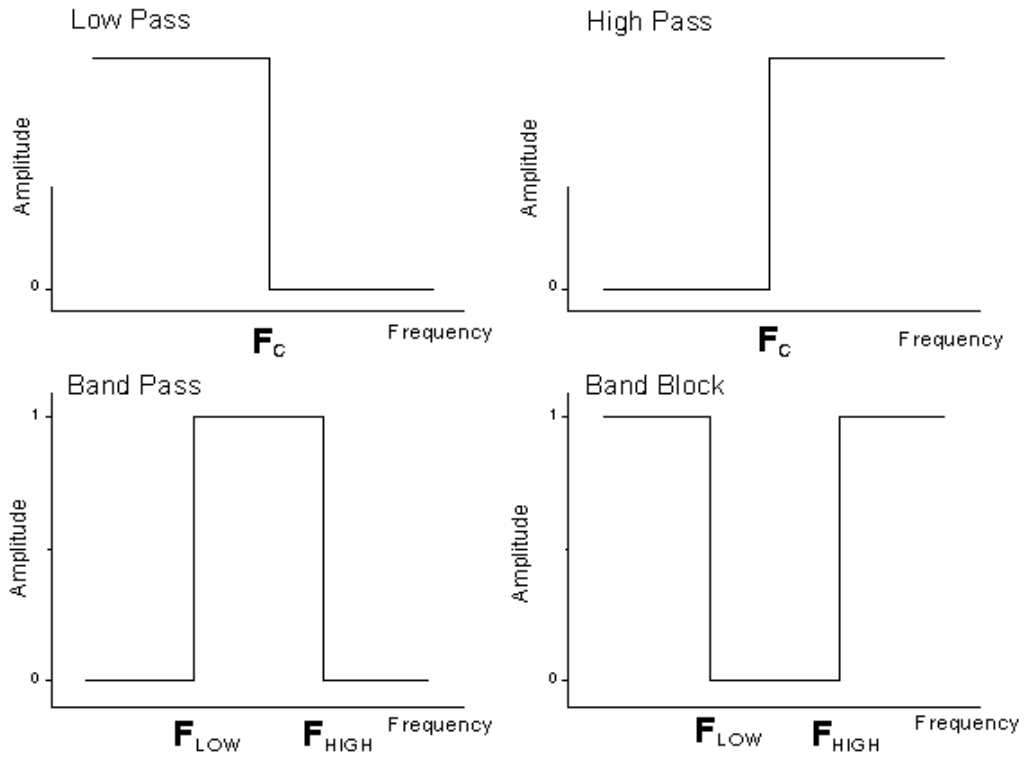
Click this button to smooth the active data plot using adjacent averaging. The Number of Points text box value controls the degree of smoothing.

The FFT Filter Button

Click this button to smooth the active data plot using FFT filter smoothing. The Number of Points text box value controls the degree of smoothing.

Digital Filtering

Origin provides five filters for digitally filtering data in the Fourier domain: low pass, high pass, band pass, band block, and threshold filters. The low and high pass filters allow you to eliminate noise at high and low frequencies, respectively. The band pass filter allows you to eliminate noise above and below a specified frequency range. The band block filter allows you to eliminate noise within a specified frequency range. The threshold filter allows you to eliminate noise corresponding to frequency components that are below a specified threshold.



Low and High Pass Filters

To apply a low or high pass filter to the active data plot, select **Analysis:FFT Filter:Low Pass** (**High Pass**). Both menu commands open the Frequency Cutoff dialog box. Origin calculates a default cutoff frequency (Fc) using the equation Fc=10*(1/period) where period is the X data set range. Type the desired cutoff frequency in the Fc text box. After clicking OK, if you selected **Low Pass**, Origin filters out frequencies above the cutoff frequency. If you selected **High Pass**, Origin filters out frequencies below the cutoff frequency. Additionally, to display the high pass filtered data with the DC offset stored in F0, select the Apply F0 Offset check box.

Origin creates a new hidden worksheet containing the X and Y components of the filtered data. Origin also displays the filtered data in the graph window.

To learn more about the low pass filter, review the FFT LOW PASS FILTER.OPJ project located in your Origin \SAMPLES\ANALYSIS\FFT folder.

Note: Origin 7 includes a number of the NAG function libraries, including c06 - Fourier Transforms. Reference information is provided on the NAG libraries in the Origin C Reference Help file (Help:Programming:Origin C Reference). Furthermore, when you install Origin, you are provided the option to install the NAG PDF files which document the NAG functions. If you clicked Yes to install these files, a \NAG PDFs folder is created with a subfolder for each library. If you did not install the

PDFs, they remain accessible on your installation CD. Furthermore, you can install them at a later date by running the Origin "Add or Remove Files" program.

To help you get started calling NAG functions from Origin C functions, a number of sample Origin projects and associated source files are included in your Origin \SAMPLES\PROGRAMMING subfolders. Those related to FFT include:

\NAG 2D FFT\

\NAG FFT Convolution\

\NAG FFT Lowpass\

\NAG STFT\

Band Pass Filter

To eliminate noise above and below a specified frequency range (band pass filter) in the active data plot, select **Analysis:FFT Filter:Band Pass**. This menu command opens the Frequency Range dialog box. Origin calculates a default low cutoff frequency (Fl) using the equation Fl=10*(1/period) where period is the X data set range. Origin calculates a default high cutoff frequency (Fh) using the equation Fh=20*(1/period). Type the desired cutoff frequencies in the Low and High text boxes. To display the filtered data with the DC offset stored in F0, select the Apply F0 Offset check box. After clicking OK, Origin filters out frequencies outside of the specified frequency range.

Origin creates a new hidden worksheet containing the X and Y components of the filtered data. Origin also displays the filtered data in the graph window.

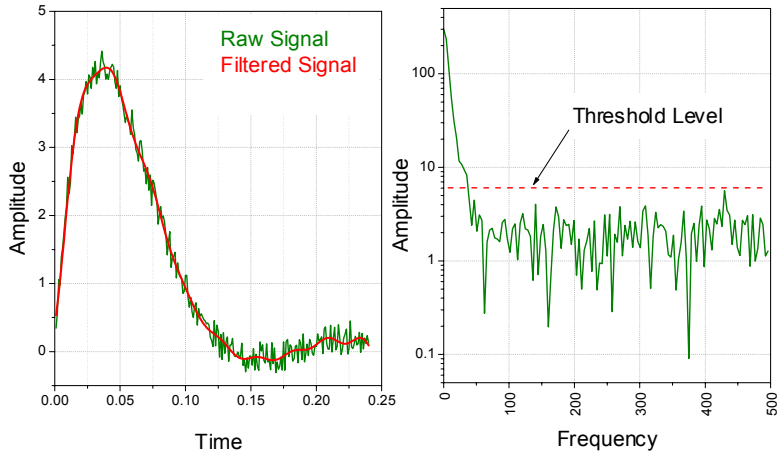
Band Block Filter

To eliminate noise within a specified frequency range (band block filter) in the active data plot, select **Analysis:FFT Filter:Band Block**. This menu command opens the Frequency Range dialog box. Origin calculates a default low cutoff frequency (Fl) using the equation Fl=10*(1/period) where period is the X data set range. Origin calculates a default high cutoff frequency (Fh) using the equation Fh=20*(1/period). Type the desired cutoff frequencies in the Low and High text boxes. To display the filtered data with the DC offset stored in F0, select the Apply F0 Offset check box. After clicking OK, Origin filters out frequencies within the specified frequency range.

Origin creates a new hidden worksheet containing the X and Y components of the filtered data. Origin also displays the filtered data in the graph window.

Threshold Filter

To eliminate noise corresponding to frequency components that are below a specified threshold level in the active data plot, select **Analysis:FFT Filter:Threshold**. This menu command first performs a forward FFT on the active data plot, and displays the frequency component spectrum with a moveable threshold level line. Drag the line to the desired level, or type an amplitude value for the threshold directly in the Threshold text box. Click the Filter Threshold button to filter out all frequency components below the set threshold level. Additionally, an inverse FFT is performed on the filtered frequency spectrum to yield a new hidden worksheet containing the original data with the noise removed.

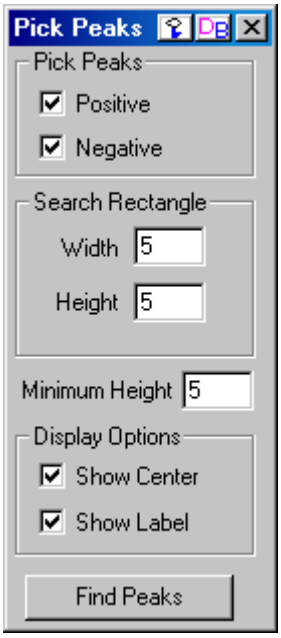


To learn more about the threshold filter, review the FFT THRESHOLD FILTER.OPJ project located in your Origin \SAMPLES\ANALYSIS\FFT folder.

Baseline and Peak Analysis

The Pick Peaks Tool

The Pick Peaks tool is available when a graph window is active. To open the tool, select **Tools:Pick Peaks**. This tool operates on the active data plot in the graph window. It enables you to find peaks when the data do not seem to have a definable baseline (for example, when the peaks appear to follow or overlap each other without intervening flat regions). The method of peak picking used for this tool does not rely on the smoothness of the data, and is, therefore, somewhat more robust than the Baseline tool.



The Pick Peaks Group

This tool can locate both positive and negative peaks in your data plot. Select the desired radio button from this group.

The Search Rectangle Group

The method for peak picking uses a moving rectangle. If a peak is to be found inside a rectangle, the difference in height between the local data maximum inside the rectangle and the data values at both rectangle ends has to be no less than the rectangle height. The height of the rectangle is the *percentage*, as specified in the Height text box, of the total amplitude of the data in the range (the amplitude is defined as the difference between the maximum and the minimum of the data). The width of the rectangle is the *percentage*, as specified in the Width text box, of the total number of points in the data range. Generally, the smaller the Height and the Width text box values, the more peaks are likely to be found. The width should not be too small, however, since the rectangle must include at least a few points.

The Minimum Height Text Box

The minimum height is the *percentage*, as specified in the Minimum Height text box, of the total amplitude of the data in the range, that the peak must have, as determined relative to the minimum of the data. The smaller the Minimum Height text box value, the more peaks are likely to be found.

The Display Options Group

Select the Show Center check box to mark the center of the plotted peak.

Select the Show Label check box to label the data plot with the X coordinate of the center of each peak.

The Find Peaks Button

Click this button to perform the peak search based on the specified settings. The peak results are stored in a hidden Peaks# worksheet, where # starts from 1 and increments each time a new data plot is processed.

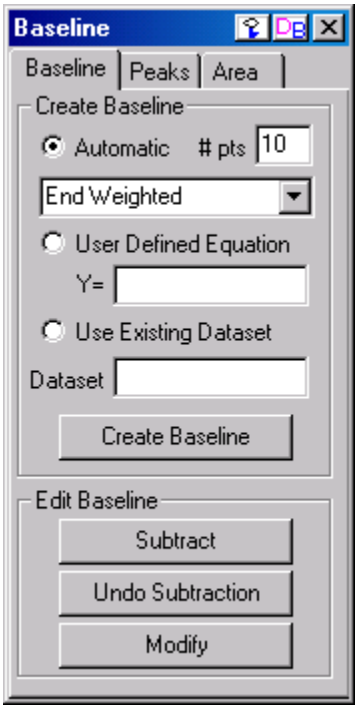
The Baseline Tool

The Baseline tool is useful for analyzing peak areas when the data plot has a definable baseline. In addition to picking peaks, it enables you to determine the extent of the peaks and the underlying baseline, as well as to integrate the data from the baseline or from the Y=0 line.

The tool consists of the Baseline tab, the Peaks tab, and the Area tab. The tool is optimally used by first finding the baseline by editing the Baseline tab, and then finding both positive and negative peaks in the data plot by editing the Peaks tab. Finally, areas under the peaks or the entire data plot can be obtained using the Area tab.

To open the Baseline tool, click on a graph window to make it active. Select **Tools:Baseline** to open the tool.

The Baseline Tab on the Baseline Tool



The Create Baseline Group

The Baseline tab provides three options for creating a baseline. After clicking the Create Baseline button, the X and Y baseline coordinates are stored in a hidden worksheet named Base#.

1) Automatically Find the Baseline

Select the Automatic radio button to find baseline points based on the baseline detection method selected from the associated drop-down list and the number of points in the # Pts text box. Click the Create Baseline button to create the baseline.

End weighted: This baseline detection method first selects data points from the two ends of the original data set. By default, the number of data points selected is approximately one-quarter of the total number of points of the original data set (one-eighth from each end). Adjacent points in this new set of data (from the ends of the original data set) are then averaged with a large window size, resulting in a smoothed data set. Baseline data points are then constructed by interpolation on the set of points common to the smoothed data set and the "end points" data set. The number of baseline points is determined by the # Pts text box value.

Entire Data w/o Smooth: Adjacent averaging is performed on the data set with a large window size, resulting in a smoothed data set. Baseline data points are then constructed by interpolation on the set of points common to the smoothed data and the original data set. The number of baseline points is determined by the # Pts text box value. This process is extremely sensitive to the distribution of the data. For instance, if the number of points on the baseline is not very large compared to the number of points that fall on the peaks, the baseline will not be well defined. In such cases, you may need to adjust the baseline points using the Modify button.

Entire Data w/ Smooth: The "Entire Data w/o Smooth" algorithm is used, but first the data is smoothed once using the Savitzky-Golay filter.

Positive Peak Algorithm: The peaks are located first, assuming there are only positive ones. The bases of each peak are then connected to form the baseline. This method is fast, but the results may not be very good in many cases.

2) User-Defined Equation

Select the User Defined Equation radio button to create a baseline for the data plot based on a user-defined equation. Type the equation in the associated text box and then click the Create Baseline button.

3) Existing Data Set

If the baseline data is available in a separate data set, you can type the name of that data set in the Dataset text box and then click the Create Baseline button to define the baseline.

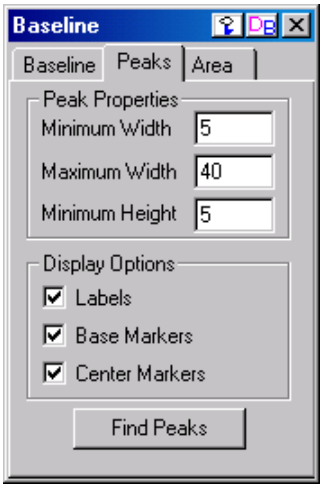
The Edit Baseline Group

Click on the Subtract button to subtract the calculated or user-specified baseline from the data plot.

Click the Undo Subtraction button to undo a baseline subtraction.

Click the Modify button to manually modify the baseline with the Data Reader tool. Click on a baseline point and drag it to a new position. If the baseline was created automatically or based on a user-defined equation, the modification to the baseline is stored in the worksheet Base#. If the baseline was created from existing data sets, the associated data sets will be modified when you move baseline points.

The Peaks Tab on the Baseline Tool



The Peak Properties Group

The minimum width of the peak is the *percentage*, as specified in the Minimum Width text box, of the total number of points in the data range that the peak must have in order to be recognized as a peak. Generally, the smaller the Minimum Width text box value, the more peaks are likely to be found.

The maximum width of the peak is the *percentage*, as specified in the Maximum Width text box, of the total number of points in the data range that the peak must not exceed in order to be recognized as a peak.

The minimum height of the peak is the *percentage*, as specified in the Minimum Height text box, of the total amplitude of the data in the range, that the peak must have (the amplitude is defined as the difference between the maximum and the minimum of the data). The smaller the Minimum Height text box value, the more peaks are likely to be found.

The Display Options Group

Select the Labels check box to label the data plot with the X coordinates of the center of each peak.

Select the Center Markers check box to mark the center of the peaks found.

Select the Base Markers check box to display markers at the boundaries (bases) of each peak.

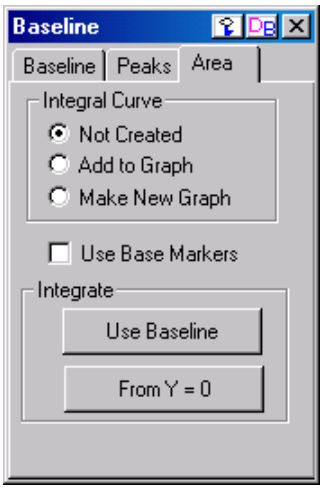
The Find Peaks Button

Click this button to find both positive and negative peaks in the current data plot. This button should be clicked after the baseline has been defined. (The baseline should be defined prior to performing this

operation.) The values under the Peak Properties group are used to determine the peak positions in the data plot. A peak is defined as a point where there is a change of sign in the first derivative, which (1) is above a user-defined distance from the baseline and (2) is flanked by a neighborhood of at least two points on each side where the first derivative is monotonic. The feet of this peak are the points where it meets the baseline.

The peak values are stored in a hidden worksheet name BsPeak#, where # starts from 1 and increments each time a new data plot is processed.

The Area Tab on the Baseline Tool



The Integral Curve Group

Select the Not Created radio button to find the area under the curve after clicking either the Use Baseline or From Y=0 buttons.

Select the Add to Graph radio button to find the area under the curve and plot the area data in the current layer after clicking either the Use Baseline or From Y=0 buttons. Rescale the Y axis of the layer to view the area data.

Select the Make New Graph radio button to find the area under the curve and plot the area data in a new default graph window after clicking either the Use Baseline or From Y=0 buttons.

The integration results are sent to the Results Log.

1) The Area column contains the area size under each peak.
2) The Center column contains the center position of the peak.
3) The Height column contains the height of the peak.

The Use Base Markers Check Box

Select this check box to include only the region between the base markers when integrating. When this check box is selected, the Add to Graph and Make New Graph radio buttons are unavailable.

The Integrate Group

Click the Use Baseline button to find the area between the active data plot and the baseline. If the Use Base Markers check box is selected, find the area between the base markers only.

Click the From Y=0 button to integrate to find the area between the active data plot and Y=0. If the Use Peak Markers check box is selected, find the area between the markers only.

Other Data Utilities

Sorting Data

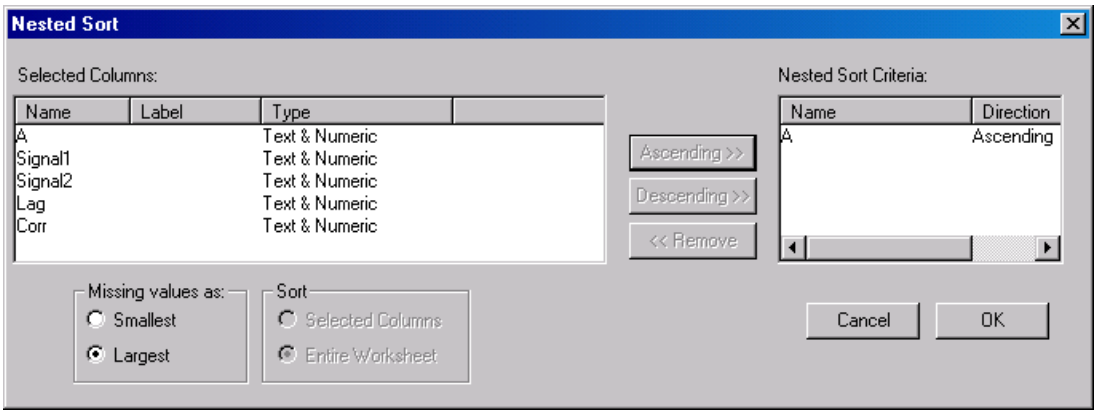
Origin can sort individual columns, multiple selected columns, a highlighted range of the worksheet, or entire worksheets. Origin offers simple sorting as well as nested sorting.

In simple sorting, the specified data is sorted using one "sort by" column and a selected sort order. To perform a simple sort of the selected data, select **Analysis:Sort Columns**. Specify **Ascending** or **Descending** from the associated submenu. If you have highlighted a range of worksheet columns, or a range of values in multiple columns, Origin sorts only the selected data based on the leftmost selected data set and the chosen sort order. If you have highlighted one column, or a range from one column, Origin sorts only the selected data. Note that if you highlight a range of rows, and not the entire column(s), the sort menu command which is enabled will be **Analysis:Sort Range**.

To perform a simple sort on the entire worksheet, select **Analysis:Sort Worksheet**. Specify **Ascending** or **Descending** from the associated submenu. Origin sorts the entire worksheet based on the leftmost selected column or the leftmost range of selected values in the worksheet and the chosen sort order. If no columns or values are selected, then Origin sorts the entire worksheet based on the leftmost column of the worksheet.

To perform a nested sort on the selected data, select **Analysis:Sort Columns:Custom** or click the Sort button on the Worksheet Data toolbar. To perform a nested sort on the entire worksheet, select **Analysis:Sort Worksheet:Custom** or select all the worksheet data and click the Sort button.

The Nested Sort Dialog Box



First select the column for the primary sort from the Selected Columns list box and click the Ascending or Descending button. Then select the column for the secondary sort from the Selected Columns list box and click the Ascending or Descending button. Select additional columns as required.

After clicking OK, Origin sorts the selected data so that the primary column is in ascending or descending order, as specified. If there are multiple rows with the same value in the primary column, the values in the corresponding rows of the secondary column, and the sort order chosen for the secondary column, are used to determine the ordering. This nesting process is continued down to the last column in the Nested Sort Criteria list box.

Normalizing Data

To normalize a data set or a range of values in a data set, highlight the desired column and select **Analysis:Normalize**. This menu command opens the Normalizing *Dataset* dialog box. The dialog box displays minima and maxima for the selected values, and provides a text box to enter a factor value. When you click OK, Origin divides all values in the selected range by the factor value.

This page is left intentionally blank.